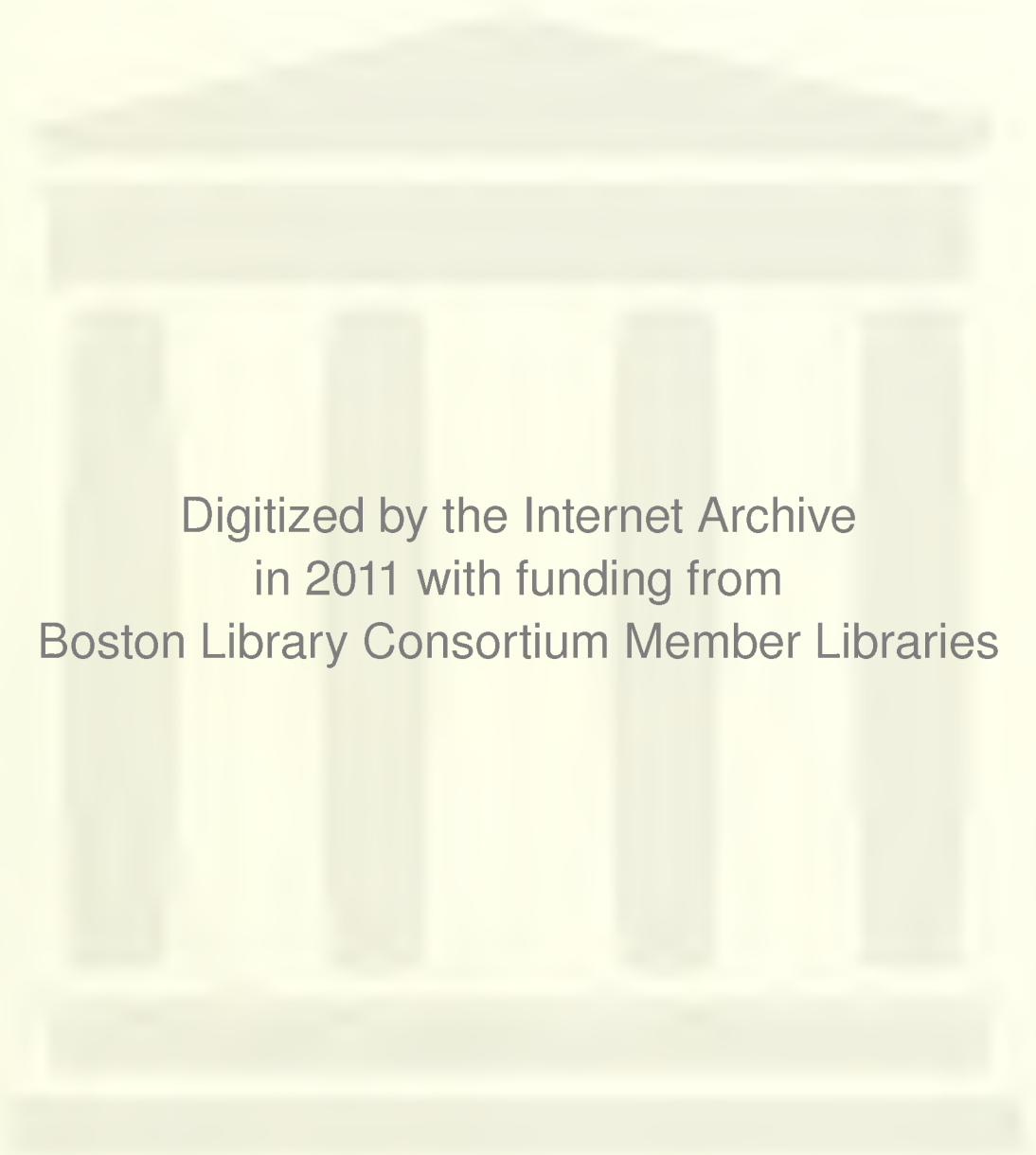


working paper
department
of economics

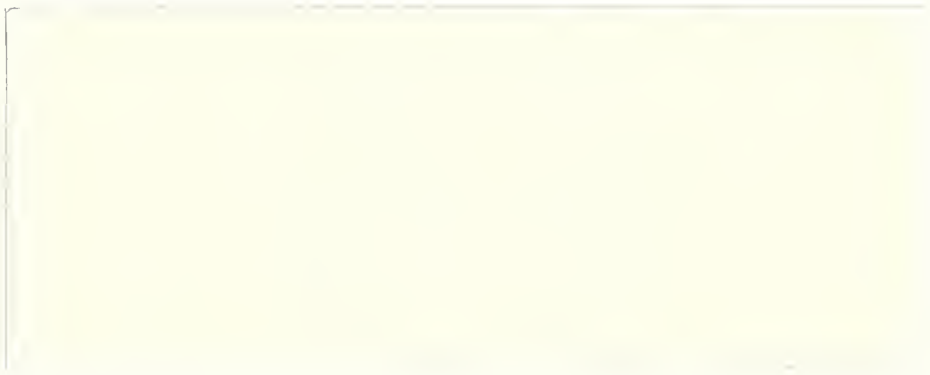
SOME ALTERNATIVES TO THE BOX-COX REGRESSION MODEL

Jeffrey M. Wooldridge

Number 534 September 1989

massachusetts
institute of
technology

50 memorial drive
cambridge, mass. 02139



# SOME ALTERNATIVES TO THE BOX-COX REGRESSION MODEL

Jeffrey M. Wooldridge

Number 534 September 1989

# SOME ALTERNATIVES TO THE BOX-COX REGRESSION MODEL

Jeffrey M. Wooldridge
Department of Economics
Massachussetts Institute of Technology, E52-262C
Cambridge, MA 02139
(617) 253-3488

March 1989
Latest Revision: September 1989

I am grateful to Ernie Berndt, Dale Poirier, and Mark Showalter for providing helpful comments. Special thanks to Ernie Berndt for directing me to the issue of scale invariance.

Abstract

A nonlinear regression model is proposed as an alternative to the Box-Cox regression model for nonnegative variables. The functional form contains as special cases the linear, exponential, constant elasticity, and generalized CES specifications, as well as other functional forms used by applied econometricians. The model can be derived from but is more general than a particular modification of the Box-Cox model. Because the model is specified directly in terms of $E(y|x)$, the parameters are easy to interpet and economic quantities are straightforward to compute. Unlike Box-Cox type approaches, the proposed weighted nonlinear least squares estimators of the conditional mean function are robust to conditional variance and other distributional misspecifications; in some leading cases they are also asymptotically efficient. Computationally simple, robust lagrange multiplier statistics for various restricted versions of the model are derived. The explained variable can be continuous, discrete, or some combination of the two. A method for obtaining scale-invariant t-statistics is also discussed, while the lagrange multiplier test for exclusion restrictions is shown to be scale invariant.

## 1. Introduction

Economists and other social scientists are often interested in explaining a nonnegative variable $y$ in terms of some explanatory variables $\mathbf{x} \equiv (x_1, x_2, \ldots, x_K)$. For many purposes this involves specifying and estimating a model for the conditional expectation of $y$ given $\mathbf{x}$. The first model encountered in econometrics courses, which postulates that $E(y|\mathbf{x})$ is a linear function of $\mathbf{x}$, or a linear function of $\phi(\mathbf{x})$ for some vector function $\phi$, often provides an inadequate description of $E(y|\mathbf{x})$. In addition, the assumption of homoskedasticity for $y$ is frequently violated by the data, resulting in the usual inference procedures being inappropriate. Finally, although of less importance for asymptotic inference, the classical assumption that $y$ conditional on $\mathbf{x}$ is normally distributed is untenable because $y$ is nonnegative.

In econometrics the most common alternative to the linear model for $E(y|\mathbf{x})$ is a linear model for $E(\log y|\mathbf{x})$, provided that $P(y > 0) = 1$. Normality of $\log y$ cannot be ruled out *a priori* and heteroskedasticity is often less of a problem in linear models with $\log y$ as the dependent variable. However, the important issue of whether the linear model for $\log y$ implicitly provides the best description of $E(y|\mathbf{x})$ depends on the particular application. This is in no way guaranteed even if the distribution of $\log y$ given $\mathbf{x}$ is normal with constant variance; one cannot even investigate this issue unless the estimates of $E(\log y|\mathbf{x})$ can be transformed into estimates of $E(y|\mathbf{x})$.

Noting that the identity and logarithmic transformations in linear models are too restrictive for all statistical applications, Box and Cox (1964) suggested a by now well-known transformation of $y$ that contains the

identity and logarithmic transformations as special cases. For nonnegative y, the Box-Cox transformation is defined as

$$y(\lambda) \equiv (y^{\lambda} - 1)/\lambda, \quad \lambda \neq 0 \tag{1.1}$$

$$\equiv \log y, \quad \lambda = 0. \tag{1.2}$$

The case $\lambda \leq 0$ is allowed only if $P(y > 0) = 1$.

In the Box-Cox regression model there is a value $\lambda \in \mathbb{R}$ such that for some Kx1 vector $\beta$ and some $\sigma^2 > 0$,

$$y(\lambda)|\mathbf{x} \sim N(\mathbf{x}\beta, \sigma^2) \tag{1.3}$$

(see also Spitzer (1982) and Hinkley and Runger (1984)). It is well known that (1.3) cannot strictly be true unless $\lambda = 0$; (1.3) should be interpreted only as an approximation. The inconsistency of the quasi-MLE's (QMLE's) of $\lambda$, $\beta$, and $\sigma^2$ due to the inherent nonnormality is well-documented (see, for example, Amemiya and Powell (1981)). Also, the practice of estimating $\lambda$ and then performing inference on $\beta$ as if $\lambda$ were known can be misleading and has been criticized by various statisticians and econometricians (see, for example, Amemiya and Powell (1981), Bickel and Doksum (1981), and Cohen and Sackrowitz (1987)).

From a social scientist's point of view there is the more important problem of interpreting the parameters $\beta$ and $\lambda$. The vector $\beta$ measures the marginal effects of the explanatory variables on $E[y(\lambda)|\mathbf{x}]$. But rarely is the variable to be explained in economic studies defined arbitrarily; the fact that y appears at all suggests that there is a natural measure of the phenomenon of interest. If y is the variable that is important to economic agents and/or policy makers then interest typically lies in the conditional expectation of y given the explanatory variables. The parameters $\beta$, $\lambda$, and $\sigma^2$ in a Box-Cox model are of interest only because they also parameterize the

conditional expectation $E(y|\mathbf{x})$. Poirier and Melino (1978) derive the relationship between $\beta$ and $E(y|\mathbf{x})$ when $y(\lambda)$ is assumed to have a plausible truncated normal distribution. They show that $\beta_j$ and $\partial E(y|\mathbf{x})/\partial x_j$ have the same sign but are not equal. But the expression for $E(y|\mathbf{x})$ depends crucially on the assumed distribution for $y(\lambda)$. In the original Box-Cox model the resulting estimates of $E(y|\mathbf{x})$ are inconsistent if in fact there is no $\lambda$ that simultaneously induces linearity of the conditional expectation, homoskedasticity, and normality. This is a potentially serious deficiency of Box-Cox type procedures since marginal effects, elasticities, and predicted values of y are of primary interest in econometric studies.

This paper offers an alternative to the Box-Cox regression model by specifying a functional form for $E(y|\mathbf{x})$ that is more flexible than simply using y or log y as the dependent variable in a linear model. The functional form analyzed here generalizes those used by others in the literature on nonlinear estimation (e.g. Mukerji (1963), Mizon (1977), Berndt and Khaled (1979)), and provides a unified framework for analyzing and testing many of the regression functions used in applied economics. As shown in section 3, it is as flexible as the Box-Cox transformation for modelling $E(y|\mathbf{x})$ but -- in contrast to Box-Cox type approaches -- tests about $E(y|\mathbf{x})$ can be carried out without imposing auxiliary distributional assumptions. The current approach is inherently more robust than models specified in terms of a nonlinear transformation of y. (For a recent example of the latter approach, see MacKinnon and McGee (1989)).

The motivation underlying this paper combines my belief that $E(y|\mathbf{x})$ and functionals of $E(y|\mathbf{x})$ are the objects of primary interest with the following observation made by Judge et. al. (1985) in their treatment of the Box-Cox

transformation:

> Despite the fact that this transformation may be useful for inducing normality on observations from skewed distributions, and despite the fact that this section appears in a chapter entitled "Nonnormal Disturbances," the main use of the Box-Cox transformation in empirical econometrics has been as a device for generalizing functional form. (p.840)

Rather than searching for a (possibly nonexistent) transformation of the explained variable that simultaneously induces approximate normality, homoskedasticity, and linearity of the conditional expectation, this paper attempts the more modest task of specifying a functional form for $E(y|\mathbf{x})$ that contains the linear, exponential, constant elasticity and a variety of other regression models as special cases. I do not worry about finding a transformation of the explained variable that is normally distributed and homoskedastic, as these features are not of primary importance for testing hypotheses in the social sciences. The parameters of the conditional mean specification are easy to interpret and the weighted nonlinear least squares estimators proposed below are likely to be sufficiently precise in many applications. In some cases the WNLS estimators are fully efficient. This notwithstanding, my view is that obtaining robust, possibly inefficient estimates of economically interesting parameters is preferred to obtaining efficient (under correct specification of the distribution) but nonrobust estimates of parameters that are difficult to interpret.

Section 2 of the paper briefly presents a case for defining all economic quantities in terms of $E(y|\mathbf{x})$, where y is the economic variable to be explained. Section 3 discusses the basic model for $E(y|\mathbf{x})$, describes how it is obtainable from a modified version of the Box-Cox model, and derives the asymptotic covariance matrix of the weighted nonlinear least squares estimator. Section 4 derives simple lagrange multiplier (LM) tests for

exclusion restrictions and for the linear and exponential special cases; both standard LM tests and LM tests that are robust to to conditional variance misspecification are covered. Section 5 extends the model to allow for Box-Cox transformations of some of the explanatory variables and discusses testing in the more general model. The important issue of obtaining scale invariant test statistics is treated in section 6. Some practical considerations are discussed in section 7, and section 8 contains concluding remarks.

## 2. Some Considerations when Choosing Functional Form

Transformations of the explained and explanatory variables are used quite liberally in the social sciences, often without regard for the implications for interpreting parameter estimates. The most common transformation for positive variables is the logarithmic transformation. If y and x are positive random scalars, a popular model is

$$E(\log y|x) = \alpha_0 + \alpha_1 \log x. \tag{2.1}$$

It follows from (2.1) that

$$\alpha_1 = \partial E(\log y|x)/\partial \log x, \tag{2.2}$$

and the coefficient $\alpha_1$ is usually interpreted as the elasticity of y with respect to x. If $\log y|x \sim N(\alpha_0+\alpha_1\log(x),\sigma^2)$ then

$$E(y|x) = \exp[\alpha_0 + \sigma^2/2 + \alpha_1 \log(x)]$$

and so

$$\alpha_1 = \partial \log E(y|x)/\partial \log x. \tag{2.3}$$

Thus, if log y conditional on x satisfies the assumptions of the classical linear model then it makes no difference whether one defines the elasticity

of y with respect to x by $\partial E(\log y|x)/\partial \log x$ or by $\partial \log E(y|x)/\partial \log x$. It is not, however, difficult to construct examples where these quantities are not the same. If in (2.1) $\log y|x \sim N(\alpha_0+\alpha_1 \log(x), 2\delta_0+2\delta_1 x)$ for some $\delta_0$, $\delta_1 > 0$ then $E(y|x) = \exp[(\alpha_0+\delta_0) + \alpha_1 \log(x) + \delta_2 x]$ and

$$\partial \log E(y|x)/\partial \log x = \alpha_1 + \delta_1 x,$$

which is always greater than $\partial E(\log y|x)/\partial x = \alpha_1$. Although this example is somewhat contrived, it is not implausible, and it does illustrate the importance of developing a unified framework in which to define economic quantities. The definitions should be as model-free as possible and the various relationships that hold, say, between derivatives and elasticities when a relationship is deterministic, should carry over to the stochastic case.

If y and x are scalars related by

$$y = f(x),$$

for a differentiable function f, then the marginal effect of x on y is simply $\partial f(x)/\partial x$, while the elasticity of y with respect to x is

$$\eta_{y,x} \equiv \frac{\partial f(x)}{\partial x} \cdot \frac{x}{f(x)}, \quad f(x) \neq 0. \tag{2.4}$$

If x, y > 0, the elasticity can also be expressed as

$$\eta_{y,x} = \frac{\partial \log f(x)}{\partial \log x}.$$

When y and x are random variables the natural definition of the marginal effect of x on y is the marginal effect of x on the <u>expected value</u> of y given x, $\partial E(y|x)/\partial x$. The advantage of defining all quantities in terms of $E(y|x)$ is that, for example, it preserves the well-known relationship (2.4) that holds between marginal effects and elasticities in the deterministic case. From the simple example above this relationship is not preserved if the

elasticity is defined in terms of E(log y|x), or in terms of some expectation other than E(y|x) (of course scaling y by a nonzero constant does not change anything). It is also straightforward to show that the deterministic relationships that hold between other economic quantities (e.g. elasticities and semi-elasticities) are preserved if they are all defined in terms of E(y|x).

This discussion extends immediately to the case of many explanatory variables. If $\mathbf{x} \equiv (x_1,\ldots,x_K)$ is a set of K explanatory variables then the marginal effect of (say) $x_K$ on E(y|x), holding $x_1,\ldots,x_{K-1}$ constant, is simply

$$\partial E(y|\mathbf{x})/\partial x_K.$$

The elasticity of y with respect to $x_K$, holding $x_1,\ldots,x_{K-1}$ constant, is

$$\eta_{y,x_K|x_1,\ldots,x_{K-1}} = \frac{\partial E(y|\mathbf{x})}{\partial x_K} \cdot \frac{x_K}{E(y|\mathbf{x})},$$

while the percentage change in E(y|x) when $x_K$ is increased by one unit is measured as

$$\frac{\partial E(y|\mathbf{x})}{\partial x_K} \cdot \frac{1}{E(y|\mathbf{x})}.$$

These measures are almost but not quite model independent. The elasticity of y with respect to $x_K$ generally changes as the list of explanatory variables changes. This is always the case in regression-type analyses and cannot be avoided, even in fully nonparametric settings.

Another benefit of defining economic quantities in terms of E(y|x) is that it circumvents the issue of whether the "disturbances" in a model for nonnegative y are additive or multiplicative. When $y \geq 0$ and interest centers on $\mu(\mathbf{x}) \equiv E(y|\mathbf{x})$, the disturbances can be multiplicative or additive:

the model

$$y = \mu(\mathbf{x}) + \epsilon, \quad E(\epsilon|\mathbf{x}) = 0 \tag{2.5}$$

is observationally equivalent to the model

$$y = \mu(\mathbf{x})e, \quad e \geq 0, \; E(e|\mathbf{x}) = 1. \tag{2.6}$$

(Simply define $\epsilon \equiv y - \mu(y|\mathbf{x})$ and $e \equiv y/\mu(\mathbf{x})$ if $P[\mu(\mathbf{x}) > 0] = 1$.) If $\epsilon$ and e are assumed to be independent of $\mathbf{x}$, then the models do differ in the conditional second moment properties of y: model (2.5) corresponds to the assumption that $V(y|\mathbf{x})$ is constant while (2.6) implies that $V(\log y|\mathbf{x})$ is constant. Although distinguishing between the two variance assumptions might be important for efficiency reasons, it is not necessary for estimating $E(y|\mathbf{x})$ or for obtaining hypotheses tests about $E(y|\mathbf{x})$ with known asymptotic size.

In summary, the point of this section is that all economic quantities should be defined in terms of $E(y|\mathbf{x})$ once the list of explanatory variables has been specified. This avoids the arbitrariness that arises if various transformations of the explained variable are entertained, and is a natural extension from the deterministic case. (One could argue that using the conditional median of y given $\mathbf{x}$ is another natural extension, but addressing this issue is beyond the scope of this paper.) The variable to be explained in most econometric studies is rarely defined arbitrarily (except possibly for the units of measurement, which only affects the scaling), so that estimates of $E[\varphi(y)|\mathbf{x}]$ for some nonlinear transformation $\varphi(y)$ are useful only if enough structure has been imposed to recover $E(y|\mathbf{x})$. Adopting the measures suggested in this section imposes a uniformity across researchers: the definition is common to all, even though different researchers might use

different functional forms for $E(y|\mathbf{x})$ and different methods of estimating $E(y|\mathbf{x})$. The common definition facilitates goodness of fit comparisions.

These points are certainly not unique to this paper; similar observations have been made in the context of specific models by Goldberger (1968), Poirier (1978), Poirier and Melino (1978), Huang and Kelingos (1979), Huang and Grawe (1980), and others. But the current paper is motivated by the observation that, in a parametric framework, the only way to estimate $E(y|\mathbf{x})$ consistently without imposing additional distributional assumptions is to specify a model for $E(y|\mathbf{x})$ directly.

## 3. Specification and Estimation of the Basic Model

Let $y$ be a nonnegative random variable and let $\mathbf{x} \equiv (x_1, x_2, \ldots, x_K)$ be a 1xK vector of explanatory variables. Typically the first element $x_1$ is unity. Without any assumptions on the conditional distribution of $y$ given $\mathbf{x}$ (except that its support is contained in $[0,\infty)$), consider the following model for $E(y|\mathbf{x})$:

$$E(y|\mathbf{x}) = [1 + \lambda \mathbf{x}\beta]^{1/\lambda}, \qquad \lambda \neq 0 \tag{3.1}$$

$$= \exp(\mathbf{x}\beta), \qquad \lambda = 0. \tag{3.2}$$

Technically, it would be more precise to replace $(\beta,\lambda)$ in (3.1) and (3.2) by something such as $(\beta_o,\lambda_o)$ to distinguish the "true" parameters from the generic parameter vector $(\beta,\lambda)$. As this results in a notational nightmare, $(\beta,\lambda)$ is used to denote the true values as well as generic values. It should be clear from the context which is intended. As usual, the vector $\mathbf{x}$ can contain nonlinear transformations of an underlying set of explanatory variables (in contrast to $y$, which should be the economic quantity of interest).

For (3.1) to be well-defined the inequality

(3.3) $$1 + \lambda x\beta \geq 0 \quad \text{(with strict inequality for } \lambda < 0)$$

must hold for all relevant values of $x$. This is analogous to requiring nonnegativity of the regression function in a Box-Cox model when $\lambda \neq 0$. Equation (3.2) is the natural definition of the regression function at $\lambda = 0$ as it is the limiting case of (3.1):

$$\lim_{\lambda \to 0} [1 + \lambda x\beta]^{1/\lambda} = \exp(x\beta).$$

Incidentally, $y$ need not be continuously distributed on $[0,\infty)$ for (3.1) and (3.2) to make sense; for example, $y$ could be a count variable.

When $\lambda = 1$ (3.1) reduces to a linear model for $E(y|x)$. The exponential regression model (3.2) is particularly appealing for nonnegative explained variables because it ensures that the predicted values are well-defined and positive for all $x$ and any value of $\beta$. Moreover, $\beta_j$ measures the constant percentage change in $E(y|x)$ when $x_j$ is increased by one unit (holding $x_1$, ..., $x_{j-1}$, $x_{j+1}$, ..., $x_K$ constant). For the general model (3.1), note that

$$\partial E(y|x)/\partial x_j = [1 + \lambda x\beta]^{((1/\lambda)-1)}\beta_j,$$

and therefore

$$\frac{\partial E(y|x)}{\partial x_j} \cdot \frac{1}{E(y|x)} = [1 + \lambda x\beta]^{-1}\beta_j.$$

The conditional mean functions (3.1) and (3.2) can be derived from a modified version of the Box-Cox model if $P(y > 0) = 1$. Recall that the Box-Cox conditional mean assumption is that for some $\beta$ and $\lambda$,

(3.4) $$E(y(\lambda)|x) = x\beta, \quad \lambda \neq 0$$

(3.5) $$E(\log y|x) = x\beta, \quad \lambda = 0.$$

Rearranging (3.4) yields

$$E(y^{\lambda}|\mathbf{x}) = 1 + \lambda\mathbf{x}\beta, \quad \lambda \neq 0. \tag{3.6}$$

Suppose now that $y|\mathbf{x}$ is lognormally distributed, so that

$$E(y|\mathbf{x}) = \exp[V(\log y|\mathbf{x})/2 + E(\log y|\mathbf{x})]. \tag{3.7}$$

Letting $m(\mathbf{x}) \equiv E(\log y|\mathbf{x})$ and $h^2(\mathbf{x}) \equiv V(\log y|\mathbf{x})$, (3.7) can be expressed as

$$m(\mathbf{x}) = \log E(y|\mathbf{x}) - h^2(\mathbf{x})/2. \tag{3.8}$$

Moreover, $w(\lambda) \equiv y^{\lambda}$ also has a lognormal distribution for $\lambda \neq 0$ with $E(\log w(\lambda)|\mathbf{x}) = \lambda m(\mathbf{x})$ and $V(\log w(\lambda)|\mathbf{x}) = \lambda^2 h^2(\mathbf{x})$. If it is assumed in addition that $h^2(\mathbf{x}) = \tau^2$ for all $\mathbf{x}$ then, for $\lambda \neq 0$,

$$\begin{aligned} \lambda m(\mathbf{x}) &= \log[E(y^{\lambda}|\mathbf{x})] - \lambda^2\tau^2/2 \\ &= \log(1 + \lambda\mathbf{x}\beta) - \lambda^2\tau^2/2 \end{aligned}$$

or

$$m(\mathbf{x}) = -\lambda\tau^2/2 + (1/\lambda)\log(1 + \lambda\mathbf{x}\beta). \tag{3.9}$$

Finally, the desired quantity $\mu(\mathbf{x}) \equiv E(y|\mathbf{x})$ can be solved for:

$$\begin{aligned} \mu(\mathbf{x}) &= \exp[\tau^2/2 + m(\mathbf{x})] \\ &= \exp[\tau^2/2 + \log(1 + \lambda\mathbf{x}\beta) - \lambda^2\tau^2/2] \\ &= \exp[(1-\lambda)\tau^2/2](1 + \lambda\mathbf{x}\beta)^{1/\lambda}, \quad \lambda \neq 0. \end{aligned} \tag{3.10}$$

The Box-Cox regression model, with the assumption of normality and homoskedasticity for log y, (rather than the assumption that $y(\lambda)$ is normally distributed and homoskedastic) yields the following functional form for $E(y|\mathbf{x})$:

$$E(y|\mathbf{x}) = \exp[(1-\lambda)\tau^2/2](1 + \lambda\mathbf{x}\beta)^{1/\lambda}, \quad \lambda \neq 0 \tag{3.11}$$

(3.12) $= \exp(\tau^2/2 + x\beta), \qquad \lambda = 0.$

Equations (3.11) and (3.12) are of the same form as (3.1) and (3.2) up to scale. However, it should be stressed that the subsequent analysis assumes only that (3.1) or (3.2) holds; no additional distributional assumptions are imposed, except those implicit in the underlying regularity conditions.

To estimate $\beta$ and $\lambda$ by nonlinear least squares (NLS) or weighted NLS (WNLS), the derivatives of the regression function are needed. Define the (K+1)x1 parameter vector $\theta \equiv (\beta', \lambda)'$ and express the parameterized regression function for $E(y|\mathbf{x})$ as

(3.13) $$\mu(\mathbf{x};\theta) \equiv [1 + \lambda x\beta]^{1/\lambda}, \qquad \lambda \neq 0$$
$$= \exp(\mathbf{x}\beta), \qquad \lambda = 0.$$

For $\lambda \neq 0$ the gradient of $\mu(\mathbf{x};\theta)$ with respect to $\beta$ is the 1xK vector

(3.14) $$\nabla_\beta \mu(\mathbf{x};\theta) = [1 + \lambda x\beta]^{\{(1/\lambda)-1\}}\mathbf{x}$$

For $\lambda = 0$,

(3.15) $$\nabla_\beta \mu(\mathbf{x};\beta,0) = \exp(\mathbf{x}\beta)\mathbf{x}.$$

Expression (3.15) is also obtained by taking the limit of (3.14):

$$\lim_{\lambda \to 0} \nabla_\beta \mu(\mathbf{x};\beta,\lambda) = \exp(\mathbf{x}\beta)\mathbf{x} = \nabla_\beta \mu(\mathbf{x};\beta,0).$$

Next consider the derivative of $\mu(\mathbf{x};\beta,\lambda)$ with respect to $\lambda$ for $\lambda \neq 0$. For $1 + \lambda c > 0$, let $q(\lambda) \equiv [1 + \lambda c]^{1/\lambda}$. Then

$$\log q(\lambda) = (1/\lambda)\log (1 + \lambda c)$$

so that

$$q'(\lambda)/q(\lambda) = [\lambda c - (1+\lambda c)\log(1+\lambda c)]/[\lambda^2(1+\lambda c)]$$

or

$$q'(\lambda) = [1 + \lambda c]^{1/\lambda}[\lambda c - (1+\lambda c)\log(1+\lambda c)]/[\lambda^2(1+\lambda c)].$$

Substituting $c \equiv x\beta$ yields

(3.16) $\nabla_\lambda \mu(\mathbf{x};\beta,\lambda) = [1+\lambda\mathbf{x}\beta]^{1/\lambda}[\lambda\mathbf{x}\beta-(1+\lambda\mathbf{x}\beta)\log(1+\lambda\mathbf{x}\beta)]/[\lambda^2(1+\lambda\mathbf{x}\beta)]$.

There are two cases of particular interest, $\lambda = 0$ and $\lambda = 1$. For $\lambda = 1$,

(3.17) $\nabla_\lambda \mu(\mathbf{x};\beta,1) = \mathbf{x}\beta - (1 + \mathbf{x}\beta)\log(1 + \mathbf{x}\beta) = \mathbf{x}\beta - E(y|\mathbf{x})\log E(y|\mathbf{x})$.

The case $\lambda = 0$ can be obtained by computing the limit of $q'(\lambda)$ as $\lambda \to 0$. Applying L'Hospital's rule twice, it can be shown that

$$\lim_{\lambda\to 0} \frac{\lambda c - (1+\lambda c)\log(1+\lambda c)}{\lambda^2(1+\lambda c)} = -c^2/2$$

so that

(3.18) $\lim_{\lambda\to 0} q'(\lambda) = -\exp(c)c^2/2.$

Thus the derivative of the regression function with respect to $\lambda$ at $\lambda = 0$ is

(3.19) $\nabla_\lambda \mu(\mathbf{x};\beta,0) = -\exp(\mathbf{x}\beta)(\mathbf{x}\beta)^2/2.$

Equation (3.17) is the basis for the LM statistic for the hypothesis $H_0$: $\lambda = 1$, while (3.19) is the basis for the LM test of $H_0$: $\lambda = 0$; both tests are developed in section 4.

Under the assumption that (3.1) or (3.2) holds $\theta$ can be consistently estimated by NLS (under (3.2), $\theta \equiv \beta$). In many cases the NLS estimator would have a large asymptotic covariance matrix due to the substantial heteroskedasticity in many nonnegative economic variables. Gains in efficiency are possible by using a weighted NLS procedure. To this end, let $\omega(\mathbf{x};\gamma) > 0$ be a weighting function that can depend on an Mx1 vector of parameters $\gamma$. The conditional mean parameters $\beta$ and $\lambda$ may be included in $\gamma$. Setting $\omega(\mathbf{x};\gamma) \equiv 1$ yields NLS. Other popular choices for $\omega$ are

(3.20) $\omega(\mathbf{x};\gamma) = [\mu(\mathbf{x};\theta)]^2 \quad (\gamma \equiv \theta)$

(3.21) $\omega(\mathbf{x};\gamma) = \mu(\mathbf{x};\theta) \quad (\gamma \equiv \theta)$

(3.22) $$\omega(\mathbf{x};\gamma) = \exp(\mathbf{x}\gamma).$$

Equation (3.20) would be appropriate if $V(y|\mathbf{x})$ is proportional to $[E(y|\mathbf{x})]^2$, while (3.21) is appropriate if $V(y|\mathbf{x})$ is proportional to $E(y|\mathbf{x})$. It must be emphasized that in what follows the weighting function is not assumed to be correctly specified for $V(y|\mathbf{x})$. In other words, it is <u>not</u> assumed that

(3.23) $$V(y|\mathbf{x}) = \sigma^2\omega(\mathbf{x};\gamma) \text{ for some } \gamma \in \mathbb{R}^M \text{ and some } \sigma^2 > 0,$$

but only that such considerations motivate the choice of $\omega$. The idea is that nonconstant weighting functions can result in efficiency gains even if (3.23) does not hold. Because the primary goal is to test hypotheses about the conditional mean parameters $\theta$, the inference should be robust to violations of (3.23).

The robust asymptotic variance-covariance matrix of the WNLS estimators of $\beta$ and $\lambda$ can be obtained by using the approach of White (1980). Let $\{(\mathbf{x}_t,y_t): t=1,2,\ldots\}$ be a sequence of random vectors following the regression model (3.1) or (3.2), and suppose there are N observations available. Assume either that these observations are independent or that they constitute a time series such that

(3.24) $$E(y_t|\mathbf{x}_t) = E(y_t|\mathbf{x}_t,\phi_{t-1}), \quad t=1,2,\ldots,$$

where $\phi_{t-1} \equiv (\mathbf{x}_{t-1},y_{t-1},\mathbf{x}_{t-2},y_{t-2},\ldots)$ denotes information observed at time t-1 ($\mathbf{x}_t$ can contain lagged dependent variables as well as other explanatory variables). Equation (3.24) ensures that the errors $\{\epsilon_t \equiv y_t - E(y_t|\mathbf{x}_t): t=1,2,\ldots\}$ are serially uncorrelated. It should be kept in mind that when $\mathbf{x}_t$ contains lagged y inequality (3.3) imposes restrictions on the distribution of $y_t$ that are generally difficult to characterize. Also, time series regressions for which (3.24) does not hold can be accomodated, but the

formulas for the asymptotic covariance matrix derived below (in particular, the formula for $B_N$) would have to be modified along the lines of White and Domowitz (1984) and Newey and West (1987).

To compute the WNLS estimator of $\theta$, estimates of the weighting functions are needed. Let $\hat{\gamma}$ denote an estimator such that $\hat{\gamma}$ would be consistent for $\gamma$ if (3.23) held. In general, plim $\hat{\gamma} = \gamma^*$ where $\gamma^*$ need not have an interpetation as "true" parameters unless (3.23) holds. If $\omega$ is chosen as in (3.20) or (3.21) then $\hat{\gamma}$ would correspond to initial estimators of $(\beta,\lambda)$, for example the NLS estimators. If $\omega$ is given by (3.22) then $\hat{\gamma}$ can be obtained by nonlinear least squares estimation using the squared NLS residual $\hat{\epsilon}_t^2$ as the regressand.

Define the weights $\hat{\omega}_t \equiv \omega(x_t;\hat{\gamma})$ (actually these are the inverse of the weights). Then the WNLS estimator $\hat{\theta}$ of $\theta$ solves

$$\min_{\theta} \sum_{t=1}^{N} (y_t - \mu(x_t;\theta))^2/\hat{\omega}_t.$$

Let $\mu_t(\theta) \equiv \mu(x_t;\theta)$, $\omega_t(\gamma) \equiv \omega(x_t;\gamma)$, $\epsilon_t \equiv y_t - \mu_t(\theta)$, $\omega_t^* \equiv \omega_t(\gamma^*)$, $\epsilon_t^* \equiv \epsilon_t/\sqrt{\omega_t^*}$, and $\nabla_\theta \mu_t^* \equiv \nabla_\theta \mu_t(\theta)/\sqrt{\omega_t^*}$. Define the (K+1)x(K+1) matrices

$$A_N \equiv N^{-1} \sum_{t=1}^{N} E[\nabla_\theta \mu_t^{*\prime} \nabla_\theta \mu_t^*]$$

$$B_N \equiv N^{-1} \sum_{t=1}^{N} E[(\epsilon_t^* \nabla_\theta \mu_t^*)' (\epsilon_t^* \nabla_\theta \mu_t^*)].$$

Then, under general heteroskedasticity of unknown form,

$$\sqrt{N}(\hat{\theta} - \theta) \xrightarrow{d} N(0, A_N^{-1} B_N A_N^{-1}).$$

If (3.23) holds then $\gamma^* = \gamma$ and the asymptotic variance of $\sqrt{N}(\hat{\theta} - \theta)$ takes the more familiar form $\sigma^2 A_N^{-1}$. In the general case a consistent estimator of $V_N \equiv A_N^{-1} B_N A_N^{-1}$ is given by the White (1980) variance-covariance matrix

estimator $\hat{V}_N \equiv \hat{A}_N^{-1}\hat{B}_N\hat{A}_N^{-1}$, where

$$\hat{A}_N \equiv N^{-1}\sum_{t=1}^{N}\nabla_\theta\hat{\mu}_t'\nabla_\theta\hat{\mu}_t/\hat{\omega}_t \equiv N^{-1}\sum_{t=1}^{N}\nabla_\theta\tilde{\mu}_t'\nabla_\theta\tilde{\mu}_t$$

$$\hat{B}_N \equiv N^{-1}\sum_{t=1}^{N}(\hat{\epsilon}_t^2/\hat{\omega}_t)(\nabla_\theta\hat{\mu}_t'\nabla_\theta\hat{\mu}_t/\hat{\omega}_t) \equiv N^{-1}\sum_{t=1}^{N}\tilde{\epsilon}_t^2\nabla_\theta\tilde{\mu}_t'\nabla_\theta\tilde{\mu}_t,$$

$\hat{\epsilon}_t \equiv y_t - \mu_t(\hat{\theta})$, $\nabla_\theta\hat{\mu}_t \equiv \nabla_\theta\mu_t(\hat{\theta})$, $\tilde{\epsilon}_t \equiv \hat{\epsilon}_t/\sqrt{\hat{\omega}_t}$, and $\nabla_\theta\tilde{\mu}_t \equiv \nabla_\theta\hat{\mu}_t/\sqrt{\hat{\omega}_t}$. The asymptotic standard error of $\hat{\theta}_j$ is the square root of the jth diagonal element of $\hat{V}_N/N$ and, at the risk of abusing the notion of convergence in distribution, one treats $\hat{\theta}$ as if $\hat{\theta} \sim N(\theta, \hat{V}_N/N)$. Carrying out inference on $\theta$ is now straightforward in cross section contexts and in time series contexts with ergodic processes and correctly specified dynamics.

Although the WNLS estimator is not always asymptotically equivalent to the maximum likelihood estimator (which is not defined in the present context), it is of course the efficient WNLS estimator if $V(y_t|\mathbf{x}_t) = \sigma^2\omega(\mathbf{x}_t;\gamma)$ and $\hat{\gamma}$ is $\sqrt{N}$-consistent for $\gamma$. In addition, there are some important cases where the WNLS estimator of $\theta$ achieves the asymptotic Cramer-Rao lower bound. If the conditional distribution of y given $\mathbf{x}$ is exponential with conditional mean $\mu(\mathbf{x};\theta)$ then the weighting function in (3.20) produces a WNLS estimator that is asymptotically equivalent to the maximum likelihood estimator. Typically the initial estimator of $\theta$ used in constructing $\hat{\omega}_t$ would be the NLS estimator. If y conditional on $\mathbf{x}$ has a Poisson distribution with mean $\mu(\mathbf{x};\theta)$ then $\omega(\mathbf{x};\gamma)$ given by (3.21) produces the WNLS estimator that is asymptotically efficient. More generally, the MLE of conditional mean parameters for densities in the linear exponential family (LEF) have asymptotically equivalent WNLS counterparts. An alternative to WNLS estimation is to simply maximize the appropriate log-likelihood function

associated with the LEF density. The expression for the standard errors derived above is still valid by letting $\hat{\omega}_t$ be the estimated variance from the distribution. For more on estimation and specification testing in these models, see Gourieroux, Monfort, and Trognon (1984) and Wooldridge (1989).

It is fairly well-known that if $\omega(\mathbf{x};\gamma)$ is misspecified for $V(y|\mathbf{x})$ then it is possible to construct a generalized method of moments (GMM) estimator that is more efficient than the WNLS estimator. This introduces additional complications into the estimation and inference procedures that are beyond the scope of this paper. If the weighting function is chosen carefully then the WNLS estimators are likely to be sufficiently precise for many applications. And if the weighting function is approximately proportional to $V(y|\mathbf{x})$ then the WNLS estimator would typically have better finite sample properties than the GMM estimator.

## 4. Lagrange Multiplier Tests for the Linear and Exponential Models

Joint estimation of $\beta$ and $\lambda$ can be difficult if the restrictions

$$1 + \hat{\lambda}\mathbf{x}_t\hat{\beta} \geq 0, \quad t=1,\ldots,N \tag{4.1}$$

have to be imposed. Before estimating the full model it makes sense to test whether some easily estimated restricted version is sufficient. The two cases of primary interest are $\lambda = 0$, which leads to an exponential regression model, and $\lambda = 1$, which is the standard linear model (without the assumptions of normality or homoskedasticity).

Before developing these tests, it is useful to briefly review the general procedure for computing the LM statistic. First consider the case where $\omega_t$ is correctly specified for $V(y_t|\mathbf{x}_t)$. Write the conditional mean function as $\mu_t(\beta,\alpha)$ such that the null hypothesis can be expressed as

$$H_0: \alpha = \alpha_0,$$

where $\beta$ is Px1 and $\alpha$ is Qx1. Let $\hat{\beta}$ be the restricted estimator of $\beta$, let $\hat{\epsilon}_t \equiv y_t - \mu_t(\hat{\beta},\alpha_0)$ be the restricted residuals, and let $\nabla_\beta \hat{\mu}_t \equiv \nabla_\beta \mu_t(\hat{\beta},\alpha_0)$ and $\nabla_\alpha \hat{\mu}_t \equiv \nabla_\alpha \mu_t(\hat{\beta},\alpha_0)$ be the gradients evaluated at the restricted estimates. In the context of WNLS quantities denoted with a "~" are the corresponding "^" variables weighted by $1/\sqrt{\hat{\omega}_t}$ (e.g. $\tilde{\epsilon}_t \equiv \hat{\epsilon}_t/\sqrt{\hat{\omega}_t}$). The usual LM statistic is $NR_u^2$ from the regression

$$\tilde{\epsilon}_t \quad \text{on} \quad \nabla_\beta \tilde{\mu}_t, \ \nabla_\alpha \tilde{\mu}_t, \quad t=1,\ldots,N, \tag{4.2}$$

where $R_u^2$ is the uncentered r-squared. Under $H_0: \alpha = \alpha_0$ <u>and</u> the assumption that $V(y_t|\mathbf{x}_t) = \sigma^2\omega(\mathbf{x}_t;\gamma)$ for some $\gamma$ and $\sigma^2$, $NR_u^2$ from (4.2) is asymptotically $\chi_Q^2$.

A form of the LM statistic that is robust to variance misspecification is not much more difficult to compute, and is originally due to Davidson and MacKinnon (1985b) for the case of unconditional heteroskedasticity and independent errors. The following procedure is taken from Wooldridge (1990) and, except that it relaxes the assumption of a correctly specified variance, it is valid under essentially the same regularity conditions needed for the nonrobust LM statistic. First, compute $\tilde{\epsilon}_t$, $\nabla_\beta \tilde{\mu}_t$, and $\nabla_\alpha \tilde{\mu}_t$ as above. Next, save the 1xQ vectors of residuals, say $\tilde{r}_t$, from the regression

$$\nabla_\alpha \tilde{\mu}_t \quad \text{on} \quad \nabla_\beta \tilde{\mu}_t, \quad t=1,\ldots,N. \tag{4.3}$$

(Note that $\tilde{r}_t$ is implicitly weighted by $1/\sqrt{\hat{\omega}_t}$). The robust LM statistic is $NR_u^2 = N - SSR$ from the regression

$$1 \quad \text{on} \quad \tilde{\epsilon}_t \tilde{r}_t, \quad t=1,\ldots,N, \tag{4.4}$$

where SSR is the sum of squared residuals. (Note that $\tilde{\epsilon}_t \tilde{r}_t$ is a 1xQ vector).

This LM statistic is asymptotically $\chi^2_Q$ under $H_0$ whether or not $\omega_t$ is correctly specified for $V(y_t|\mathbf{x}_t)$, and it is asymptotically equivalent to the nonrobust LM test if $\omega_t$ is correctly specified for $V(y_t|\mathbf{x}_t)$.

Consider now testing the null hypothesis $H_0$: $\lambda = 1$ in the model (3.1). Let $\delta \equiv (1+\beta_1,\beta_2,\beta_3,\ldots,\beta_K)$ and assume that $x_1 \equiv 1$ for expository purposes. Under $H_0$ (3.1) reduces to $E(y|\mathbf{x}) = \mathbf{x}\delta$. Expressions (3.14) and (3.17) provide the gradient of the regression function evaluated at the null hypothesis:

$$\nabla_\theta \mu(\mathbf{x};\beta,1) = (\mathbf{x},\ \mathbf{x}\beta - (1 + \mathbf{x}\beta)\log(1 + \mathbf{x}\beta)). \tag{4.5}$$

Let $\hat{\delta}$ be the WLS estimator from the regression

$$y_t \text{ on } 1,\ x_{t2},\ \ldots,\ x_{tK},\ t=1,\ldots,N \tag{4.6}$$

using weights $1/\sqrt{\hat{\omega}_t}$, and let the (unweighted) predicted values and residuals be $\hat{y}_t \equiv \mathbf{x}_t\hat{\delta}$, $\hat{\epsilon}_t \equiv y_t - \hat{y}_t$. The weighted residuals are $\tilde{\epsilon}_t \equiv \hat{\epsilon}_t/\sqrt{\hat{\omega}_t}$. Because $\sum_{t=1}^{N} \mathbf{x}_t'\hat{\epsilon}_t/\hat{\omega}_t = 0$ and $1 + \mathbf{x}\beta = \mathbf{x}\delta$, the usual LM statistic is obtained as $NR_u^2$ from the regression

$$\tilde{\epsilon}_t \text{ on } \tilde{\mathbf{x}}_t,\ \tilde{y}_t\log(\hat{y}_t) \quad t=1,2,\ldots,N,$$

where $\tilde{\mathbf{x}}_t \equiv \mathbf{x}_t/\sqrt{\hat{\omega}_t}$ and $\tilde{y}_t \equiv \hat{y}_t/\sqrt{\hat{\omega}_t}$. Under $H_0$ and $V(y_t|\mathbf{x}_t) = \sigma^2\omega_t(\gamma)$, $NR_u^2 \overset{a}{\sim} \chi^2_1$. When $\hat{\omega}_t \equiv 1$ this reduces to the LM form of the t-test for the null of a linear model against the Box-Cox alternative derived by Andrews (1971). Because Andrews starts from the Box-Cox model, his derivation of the statistic is different. Andrews suggests using the standard t-test in the Box-Cox setup because, under $H_0$: $\lambda = 1$, $y|\mathbf{x}$ is normally distributed and homoskedastic.

The failure of normality has no effect on the asymptotic size of the test, but misspecification of the conditional variance function can bias the

inference toward or away from the linear conditional expectation. Because the primary goal is to test hypotheses about $E(y|\mathbf{x})$ it is prudent to use the LM test that is robust to misspecification of $V(y|\mathbf{x})$. Let $\tilde{r}_t$ be the residuals from the regression

$$\tilde{y}_t \log(\hat{y}_t) \quad \text{on} \quad \tilde{\mathbf{x}}_t, \quad t=1,2,\ldots N.$$

Run the regression

$$1 \quad \text{on} \quad \tilde{\epsilon}_t \tilde{r}_t, \quad t=1,2,\ldots,N$$

and use N - SSR as asymptotically $\chi^2_1$ under $H_0$.

If $\hat{y}_t$ is nonpositive for some t then the indicator $\hat{y}_t \log(\hat{y}_t)$ is not defined. This may suggest that observation t is an outlier; an alternative interpretation is that the hypothesis $\lambda = 1$ is false.

As Davidson and MacKinnon (1985a) have emphasized Andrew's test (based on ordinary least squares rather than weighted least squares) is not optimal if the standard Box-Cox model holds (it ignores the changing shape of the distribution of $y|\mathbf{x}$ as $\lambda$ varies). But the optimal test is not robust to violations of the normality and homoskedasticity assumptions for $y(\lambda)$. The tests suggested by Davidson and MacKinnon (1985a) are joint tests of the distributional and conditional mean assumptions imposed in the Box-Cox model. The LM test for $\lambda = 1$ based on a more plausible truncated normal distribution is difficult to compute and not robust to failure of the truncated normal distributional assumption (see Poirier and Ruud (1979,1983)). Seaks and Layson (1983) provide strong evidence that heteroskedasticity of $y(\lambda)$ in the Box-Cox model can seriously bias estimates and test statistics. On the other hand, the robust form of Andrew's test is easy to compute, maintains the correct asymptotic size under misspecification of $V(y|\mathbf{x})$, and is likely to have sufficient power for many applications. Also, the robust form of the

test is asymptotically equivalent to the usual LM test if $\omega_t$ happens to be correctly specified for $V(y_t|\mathbf{x}_t)$.

The LM test for $\lambda = 0$ simply requires weighted nonlinear least squares estimation of an exponential regression function, which is relatively straightforward. Let $\hat{\beta}$ be the WNLS estimator of $\beta$ in the model

$$E(y_t|\mathbf{x}_t) = \exp(\mathbf{x}_t\beta) \tag{4.7}$$

using weights $\hat{\omega}_t$. Thus, $\hat{\beta}$ solves

$$\min_{\beta} \sum_{t=1}^{N} (y_t - \exp(\mathbf{x}_t\beta))^2/\hat{\omega}_t.$$

Let $\hat{y}_t \equiv \exp(\mathbf{x}_t\hat{\beta})$ be the fitted values, and let $\hat{\epsilon}_t \equiv y_t - \hat{y}_t$ be the residuals from the WNLS estimation. Then, refering to (3.19), the LM statistic is based on the scalar

$$\sum_{t=1}^{N} \exp(\mathbf{x}_t\hat{\beta})(\mathbf{x}_t\hat{\beta})^2\hat{\epsilon}_t/\hat{\omega}_t \tag{4.8}$$

or

$$\sum_{t=1}^{N} \tilde{y}_t(\log \hat{y}_t)^2\tilde{\epsilon}_t. \tag{4.9}$$

The LM test for $\lambda = 0$ if $V(y|\mathbf{x}) = \sigma^2\omega(\mathbf{x};\gamma)$ is obtained as $NR_u^2$ from the regression

$$\tilde{\epsilon}_t \text{ on } \tilde{y}_t\mathbf{x}_t,\ \tilde{y}_t(\log \hat{y}_t)^2, \quad t=1,\ldots,N.$$

The robust form of the test can be computed by first saving the residuals $\tilde{r}_t$ from the regression

$$\tilde{y}_t(\log \hat{y}_t)^2 \text{ on } \tilde{y}_t\mathbf{x}_t, \quad t=1,\ldots,N$$

and then calculating N - SSR from the regression

$$1 \text{ on } \tilde{\epsilon}_t\tilde{r}_t, \quad t=1,2,\ldots,N;$$

N - SSR is asymptotically $\chi^2_1$ under $H_0$.

The LM test for exclusion restrictions is also easy to derive. Let $z_t$ be a 1xQ vector of additional variables, and consider testing $H_0$: $\delta = 0$ in the model

$$E(y_t|x_t,z_t) = [1 + \lambda x_t\beta + \lambda z_t\delta]^{1/\lambda} \tag{4.10}$$

(note that $E(y_t|x_t,z_t) = E(y_t|x_t)$ under $H_0$). Let $\hat{\beta}$ and $\hat{\lambda}$ be the estimates computed under $\delta = 0$, so that the fitted values and residuals computed under $H_0$ are $\hat{y}_t \equiv [1 + \hat{\lambda}x_t\hat{\beta}]^{1/\hat{\lambda}}$ and $\hat{\epsilon}_t \equiv y_t - \hat{y}_t$, respectively. Let $\nabla_\beta\hat{\mu}_t$ and $\nabla_\lambda\hat{\mu}_t$ be the gradients defined by (3.15) and (3.17). The gradient with respect to $\delta$ evaluated under the null is

$$\nabla_\delta\hat{\mu}_t \equiv [1 + \hat{\lambda}x_t\hat{\beta}]^{\{(1/\hat{\lambda}) - 1\}}z_t.$$

If WNLS is used, each quantity is weighted by $1/\sqrt{\hat{\omega}_t}$ and the LM test is obtained as $NR_u^2$ from the regression

$$\tilde{\epsilon}_t \quad \text{on} \quad \nabla_\beta\tilde{\mu}_t,\ \nabla_\lambda\tilde{\mu}_t,\ \nabla_\delta\tilde{\mu}_t; \tag{4.11}$$

the robust LM test first requires the 1xQ residuals $\tilde{r}_t$ from

$$\nabla_\delta\tilde{\mu}_t \quad \text{on} \quad \nabla_\beta\tilde{\mu}_t,\ \nabla_\lambda\tilde{\mu}_t, \tag{4.12}$$

and using $\tilde{r}_t$ as in (4.4). Exclusion restriction tests when $\lambda$ is fixed at $\lambda = 0$ or $\lambda = 1$ are even easier to compute.

It is important to emphasize that tests about $\beta$ and $\lambda$ (and $\delta$ in (4.10)) in the current setup are purely tests about $E(y|\mathbf{x})$: provided the robust forms of the tests are used, the null hypothesis imposes no assumptions on $V(y|\mathbf{x})$ or any other aspect of the distribution of y given $\mathbf{x}$. In contrast, tests about $\beta$ and $\lambda$ in the Box-Cox model are generally tests about the <u>entire</u> distribution of y given $\mathbf{x}$. It is quite conceivable that one could reject the

null hypothesis because the distribution is misspecified even if the conditional mean is (implicitly) correctly specified. The efficiency of the tests based on Box-Cox type approaches comes at a substantial price. Not only are the tests nonrobust, but it is not possible to isolate the hypothesis that $E(y|\mathbf{x})$ is correctly specified.

## 5. A Model with Box-Cox Transformations of the Explanatory Variables

The model presented in section 3 can easily incorporate Box-Cox transformations of the explanatory variables (which is <u>not</u> qualitatively the same as transforming y). Suppose that a Box-Cox transformation is to be applied to the nonnegative variables $x_j$, $j=J+1,\ldots,K$. Then (3.1) can be extended to

$$\mu(\mathbf{x};\theta) \equiv [1 + \lambda\{\beta_1 x_1+\ldots+\beta_J x_J+\beta_{J+1}x_{J+1}(\rho_{J+1})+\ldots+\beta_K x_K(\rho_K)\}]^{1/\lambda}, \quad \lambda \neq 0 \tag{5.1}$$

$$\equiv \exp[\beta_1 x_1+\ldots+\beta_J x_J+\beta_{J+1}x_{J+1}(\rho_{J+1})+\ldots+\beta_K x_K(\rho_K)], \quad \lambda = 0. \tag{5.2}$$

Here, $x_j(\rho_j)$ denotes the Box-Cox transformation of $x_j$ with parameter $\rho_j$, as in (1.1) and (1.2). Equations (5.1) and (5.2) significantly expand the range of nested special cases. If $\lambda = 0$ and $\rho_j = 0$, $\partial \log E(y|\mathbf{x})/\partial \log(x_j) = \beta_j$, so that $\beta_j$ is the elasticity of y with respect to $x_j$ (as defined in section 2). If $\lambda = 1$ and $\rho_j = 0$ then $\beta_j = \partial E(y|\mathbf{x})/\partial \log(x_j)$, so that $\beta_j$ measures the change in the expected value of y when $x_j$ increases by one percent.

Equation (5.1) also contains the CES production function as a special case. If $x_1 \equiv 1$ and $x_2,,\ldots,x_K$ are nonnegative inputs (take J=1), the unrestricted form of (5.1) is

$$[1 + \lambda\beta_1 + \lambda\beta_2 x_2(\rho_2) + \ldots + \lambda\beta_K x_K(\rho_K)]^{1/\lambda}$$

$$\equiv [\alpha_1 + \alpha_2 x_2(\rho_2) + \ldots + \alpha_K x_K(\rho_K)]^{1/\lambda}.$$

The CES function is obtained under the restriction $\lambda = \rho_2 = \ldots = \rho_K$.

The exponential form (5.2) has the advantage of ensuring that the predicted values are well-defined and positive without any restrictions on the parameters (except $\rho_j > 0$ if $P(x_j = 0) > 0$). Such a model is reasonably easy to estimate. Also, (5.2) includes the constant elasticity and constant semi-elasticity cases as restricted versions. These models have been fairly popular in applied econometric studies, particularly in models of count data (see Hausman, Hall, and Griliches (1984) and Papke (1989)). The LM test for $\lambda = 0$ developed in section 4 and the extensions discussed below might be useful specification tests of the exponential model with $\hat{\omega}_t \equiv \exp(x_t\hat{\beta})$.

The derivatives of $\mu(\mathbf{x};\theta)$ with respect to the parameters $\beta$ and $\lambda$ are similar to those already obtained, with the exception that $x_{J+1},\ldots,x_K$ are replaced by $x_{J+1}(\rho_{J+1}),\ldots,x_K(\rho_K)$. Let $\mathbf{x}(\rho)$ denote the 1xK vector

$$\mathbf{x}(\rho) \equiv (x_1,\ldots,x_J,x_{J+1}(\rho_{J+1}),\ldots,x_K(\rho_K))$$

so that

$$\mu(\mathbf{x};\theta) \equiv [1 + \lambda\mathbf{x}(\rho)\beta]^{1/\lambda}, \qquad \lambda \neq 0$$

$$\mu(\mathbf{x};\theta) \equiv \exp[\mathbf{x}(\rho)\beta], \qquad \lambda = 0.$$

Then, by (3.14) and (3.16), for $\lambda \neq 0$

$$\nabla_\beta \mu(\mathbf{x};\beta,\lambda,\rho) = [1 + \lambda\mathbf{x}(\rho)\beta]^{\{(1/\lambda)-1\}}\mathbf{x}(\rho) \tag{5.3}$$

$$\nabla_\lambda \mu(\mathbf{x};\beta,\lambda,\rho) = [1+\lambda\mathbf{x}(\rho)\beta]^{1/\lambda}[\lambda\mathbf{x}(\rho)\beta-(1+\lambda\mathbf{x}(\rho)\beta)\log(1+\lambda\mathbf{x}(\rho)\beta)]/[\lambda^2(1+\lambda\mathbf{x}(\rho)\beta)]. \tag{5.4}$$

To obtain the derivative with respect to $\rho_j$, j=J+1,...,K, note that if $z(\rho)$ denotes the Box-Cox transformation of z, then

$$\partial z(\rho)/\partial\rho = z(\rho)\log(z) - (z(\rho) - \log(z))/\rho, \qquad \rho \neq 0$$

$$= (\log z)^2/2 \qquad \rho = 0.$$

Thus, for $\lambda \neq 0$, if $\nabla_\rho \mathbf{x}(\rho)$ denotes the KxK gradient of $\mathbf{x}(\rho)$ with respect to $\rho \equiv (\rho_{J+1},\ldots,\rho_K)$ then the 1x(K-J) gradient of $\mu$ with respect to $\rho$ is

$$\nabla_\rho \mu(\mathbf{x};\beta,\lambda,\rho) = [1 + \lambda\mathbf{x}(\rho)\beta]^{\{(1/\lambda)-1\}}\beta'\nabla_\rho \mathbf{x}(\rho)'. \tag{5.5}$$

The first J columns of $\nabla_\rho \mathbf{x}(\rho)$ are zero while the only nonzero element in the (J+i)th column is in the ith row; this term is equal to

$$x_{J+i}(\rho_{J+i})\log(x_{J+i}) - (x_{J+i}(\rho_{J+i}) - \log(x_{J+i}))/\rho_{J+i}$$

or, if $\rho_{J+i} = 0$, $(\log x_{J+i})^2/2$, i=1,...,K-J.

The derivatives for $\lambda = 0$ are easily seen to be

$$\nabla_\beta \mu(\mathbf{x};\beta,0,\rho) = \exp[\mathbf{x}(\rho)\beta]\mathbf{x}(\rho) \tag{5.6}$$

$$\nabla_\lambda \mu(\mathbf{x};\beta,0,\rho) = -\exp[\mathbf{x}(\rho)\beta][\mathbf{x}(\rho)\beta]^2/2. \tag{5.7}$$

$$\nabla_\rho \mu(\mathbf{x};\beta,0,\rho) = \exp[\mathbf{x}(\rho)\beta]\beta'\nabla_\rho \mathbf{x}(\rho)'. \tag{5.8}$$

The K+1+(K-J) parameter vector $\theta \equiv (\beta',\lambda,\rho')'$ can be estimated by WNLS as in section 3. The asymptotic variance and its estimator derived there are still valid once the gradient is redefined to be

$$\nabla_\theta \mu_t(\theta) \equiv [\nabla_\beta \mu_t(\beta,\lambda,\rho), \nabla_\lambda \mu_t(\beta,\lambda,\rho), \nabla_\rho \mu_t(\beta,\lambda,\rho)].$$

The asymptotic variance of any restricted version is obtained by calculating the gradient of $\mu_t(\theta)$ with respect to the unrestricted elements. Two cases of particular interest are $\lambda = 1$ and $\lambda = 0$. In these cases $\theta$ contains only

$(\beta,\rho)$ and the gradients are given by (5.3) and (5.5) ($\lambda$=1) or (5.6) and (5.8) ($\lambda$=0).

Several models used by applied economists are special cases of (5.1) or (5.2). The null of a linear model is stated as

$$H_0\colon\ \lambda = 1,\ \rho_j = 1,\ j=J+1,\ldots,K,$$

and the constrained derivatives are

$$\nabla_\beta \mu(\mathbf{x};\beta,1,1) = \mathbf{x}(1) = (x_1,\ldots,x_J,x_{J+1}-1,\ldots,x_K-1)$$

$$\nabla_\lambda \mu(\mathbf{x};\beta,1,1) = \mathbf{x}(1)\beta - (1 + \mathbf{x}(1)\beta)\log(1 + \mathbf{x}(1)\beta)$$

$$\nabla_\rho \mu(\mathbf{x};\beta,1,1) = \beta' \nabla_\rho \mathbf{x}(1)'$$

$$= (\beta_{J+1}[x_{J+1}\log(x_{J+1}) - (x_{J+1}-1)],\ldots,\beta_K[x_K\log(x_K) - (x_K-1)]).$$

Suppose that x contains unity and let $\hat{\delta}$ denote the WLS estimator from the regression

$$y_t \text{ on } \mathbf{x}_t, \quad t=1,\ldots,N, \quad \text{using weights } \hat{\omega}_t.$$

Define $\hat{y}_t \equiv \mathbf{x}_t\hat{\delta}$, $\hat{\epsilon}_t \equiv y_t - \hat{y}_t$. Because $\sum_{t=1}^{N} \mathbf{x}_t'\hat{\epsilon}_t/\hat{\omega}_t = 0$, the LM test is based on the K-J+1 sample covariances

$$\sum_{t=1}^{N} \hat{y}_t\log(\hat{y}_t)\hat{\epsilon}_t/\hat{\omega}_t \equiv \sum_{t=1}^{N} \tilde{y}_t\log(\hat{y}_t)\tilde{\epsilon}_t$$

$$\sum_{t=1}^{N} x_{tj}\log(x_{tj})\hat{\epsilon}_t/\hat{\omega}_t \equiv \sum_{t=1}^{N} \tilde{x}_{tj}\log(x_{tj})\tilde{\epsilon}_t, \quad j=J+1,\ldots,K$$

where quantities with a "~" are the corresponding "^" quantities weighted by $1/\sqrt{\hat{\omega}_t}$. If $V(y_t|\mathbf{x}_t) = \sigma^2\omega_t(\gamma)$ under $H_0$ then the LM test is simply $NR_u^2$ from the regression

$$\tilde{\epsilon}_t \text{ on } \tilde{\mathbf{x}}_t,\ \tilde{y}_t\log(\hat{y}_t),\ \tilde{x}_{t,J+1}\log(x_{t,J+1}),\ \ldots,\ \tilde{x}_{tK}\log(x_{tK});$$

$NR_u^2 \sim \chi^2_{K-J+1}$ under $H_0$. It is probably best to use the form that is robust to second moment misspecification. Define the 1x(K-J+1) vector

$$\hat{\xi}_t \equiv (\hat{y}_t \log(\hat{y}_t),\ x_{t,J+1}\log(x_{t,J+1}),\ \ldots,\ x_{tK}\log(x_{tK})),$$

$\tilde{\xi}_t \equiv \hat{\xi}_t/\sqrt{\hat{\omega}_t}$, and $\tilde{x}_t \equiv x_t/\sqrt{\hat{\omega}_t}$. Then regress

$$\tilde{\xi}_t \quad \text{on} \quad \tilde{x}_t$$

and save the K-J+1 residuals, say $\tilde{r}_t$. Then $NR_u^2 = N - SSR$ from the regression

$$1 \quad \text{on} \quad \tilde{\epsilon}_t\tilde{r}_t, \quad t=1,\ldots,N \tag{5.9}$$

is asymptotically $\chi^2_{K-J+1}$ under $H_0$ (which does not impose $V(y|x) = \sigma^2\omega(x;\gamma)$).

If the linear model is rejected one might test the less restrictive hypothesis $H_0$: $\lambda = 1$. This is a one degree of freedom test with misspecification indicator

$$\hat{\xi}_t \equiv \nabla_\lambda \mu(x_t;\hat{\beta},1,\hat{\rho}) =$$

$$x_t(\hat{\rho})\hat{\beta} - (1 + x_t(\hat{\rho})\hat{\beta})\log(1 + x_t(\hat{\rho})\hat{\beta}),$$

where $\hat{\beta}$ and $\hat{\rho}$ now denote the estimators computed under the single restriction $\lambda = 1$. Let $\hat{y}_t \equiv 1 + x_t(\hat{\rho})\delta$, $\hat{\epsilon}_t \equiv y_t - \hat{y}_t$, $\hat{x}_t \equiv x_t(\hat{\rho})$, and $\nabla_\rho \hat{x}_t \equiv (\nabla_{\rho_{J+1}}\hat{x}_{t,J+1},\ldots,\nabla_{\rho_K}\hat{x}_{tK})$ (a 1x(K-J) vector). Then the usual LM test involves regressing

$$\tilde{\epsilon}_t \quad \text{on} \quad \tilde{x}_t,\ \nabla_\rho\tilde{x}_t,\ \tilde{y}_t\log(\hat{y}_t)$$

and using $NR_u^2$ as $\chi^2_1$. The robust form uses N-SSR from the regression

$$1 \quad \text{on} \quad \tilde{\epsilon}_t\tilde{r}_t,$$

where $\tilde{r}_t$ are the scalar residuals from the regression

$$\tilde{y}_t\log(\hat{y}_t) \quad \text{on} \quad \tilde{x}_t,\ \nabla_\rho\tilde{x}_t.$$

Under $\lambda = 0$, $\rho_j = 0$, $j=J+1,\ldots,K$, the model reduces to

$$E(y|x) = \exp[\beta_1 x_1 + \ldots + \beta_J x_J + \beta_{J+1}\log(x_{J+1}) + \ldots + \beta_K \log(x_K)],$$

so that $\beta_j$ for $j \geq J+1$ is an elasticity while $\beta_j$ for $j \leq J$ measures the percentage change in $E(y|x)$ when $x_j$ increases by one unit. It is traditional to estimate these quantities from the regression

$$\log(y_t) \quad \text{on} \quad x_{t1}, \ldots, x_{tJ}, \log(x_{t,J+1}), \ldots, \log(x_{tK})$$

but, as pointed out in section 2, the two procedures need not give the same answer, even asymptotically. Let $\hat{\beta}$ denote the WNLS estimator of $\beta$ from the weighted nonlinear regression

$$y_t \quad \text{on} \quad \exp[\beta_1+\ldots+\beta_J x_{tJ}+\beta_{J+1}\log(x_{t,J+1})+\ldots+\beta_K \log(x_{tK})]$$

using weights $\hat{\omega}_t$, and let $\hat{y}_t$ and $\hat{\epsilon}_t$ be the (unweighted) fitted values and residuals ($x_{t1}$ has been set to unity). Evaluating the gradient of $\mu(\mathbf{x}_t;\beta,\lambda,\rho)$ at $(\hat{\beta},0,0)$ and weighting all quantities by $1/\sqrt{\hat{\omega}_t}$ leads to the auxiliary regression

$$\tilde{\epsilon}_t \quad \text{on} \quad \tilde{y}_t\mathbf{x}_t(0), \tilde{y}_t(\log \hat{y}_t)^2, \tilde{y}_t(\log x_{t,J+1})^2, \ldots, \tilde{y}_t(\log x_{tK})^2$$

where $\mathbf{x}_t(0) \equiv (1,x_{t2},\ldots,x_{tJ},\log(x_{t,J+1}),\ldots,\log(x_{tK}))$. The LM statistic $NR_u^2$ from this regression is $\chi^2_{K-J+1}$ under $H_0$ if $V(y|\mathbf{x}) = \sigma^2\omega(\mathbf{x};\gamma)$. The robust test is obtained by defining

$$\tilde{\xi}_t \equiv (\tilde{y}_t(\log \hat{y}_t)^2,\tilde{y}_t(\log x_{t,J+1})^2,\ldots,\tilde{y}_t(\log x_{tK})^2),$$

regressing $\tilde{\xi}_t$ on $\tilde{y}_t\mathbf{x}_t(0)$ and saving the residuals $\tilde{r}_t$, and using $\tilde{\epsilon}_t$ and $\tilde{r}_t$ as in (5.9).

The test for $H_0$: $\lambda = 0$ uses the scalar weighted indicator $\tilde{\xi}_t \equiv \tilde{y}_t(\log \hat{y}_t)$, where the fitted values are now $\hat{y}_t \equiv \exp[\mathbf{x}_t(\hat{\rho})\hat{\beta}]$, $\hat{\beta}$ and $\hat{\rho}$ are now computed from WNLS of (5.2), and $\tilde{y}_t \equiv \hat{y}_t/\sqrt{\hat{\omega}_t}$. As usual, the residuals are $\hat{\epsilon}_t = y_t - \hat{y}_t$ and $\tilde{\epsilon}_t = \hat{\epsilon}_t/\sqrt{\hat{\omega}_t}$. Let $\tilde{r}_t$ be the residuals from the regression

$$\tilde{\xi}_t \quad \text{on} \quad \tilde{y}_t\hat{x}_t,\ \tilde{y}_t\nabla_\rho\hat{x}_t,$$

and then use N - SSR from 1 on $\tilde{\epsilon}_t\tilde{r}_t$ as $\chi^2_1$ under $H_0$.

LM tests of other restrictions can be obtained by computing the residuals and gradients under the null hypothesis and following procedures analagous to those outlined above. The hypothesis $H_0$: $\lambda = 0$, $\rho_j = 1$, j=J+1,...,K is likely to be of general interest; $H_0$: $\lambda = \rho_2 = \ldots = \rho_K$ is of interest in the CES example. Testing the exclusion restrictions $H_0$: $\delta = 0$ in the model

$$E(y_t|x_t) = [1 + \lambda x_t(\rho)\beta + \lambda z_t\delta]^{1/\lambda},$$

where $z_t$ is 1xQ, is similar to the case covered at the end of section 4, except that the gradient $\nabla_\rho\hat{\mu}_t$ (see (5.5)) must be included in (4.11) (nonrobust test) or (4.12) (robust test). Note that the variables $z_t$ cannot themselves be transformed because the transformation parameters are not identified under $\delta = 0$.

## 6. On the Issue of Scale Invariance

One feature of the model (3.1) (and the more general model (5.1)) might make some researchers uncomfortable: the t-statistics for the slope coefficients $\hat{\beta}_2,\ldots,\hat{\beta}_K$ are not invariant to the scaling of $y_t$. This is in contrast to the case of linear regression or exponential regression, where, for example, it does not matter for purposes of inference whether y is measured in hundreds or thousands of dollars. In the linear case the coefficients are scaled up or down but the t-statistics are unchanged. For the exponential regression model (3.2) it is easy to see that only the constant term $\beta_1$ changes when y is scaled, and the standard errors of all coefficients are invariant. Unfortunately, this does not carry over to the

general model (3.1) when $\lambda$ is estimated along with $\beta$. Spitzer (1984) pointed out the analogous feature for the standard Box-Cox model.

Focusing on (3.1), suppose that $(\hat{\beta},\hat{\lambda})$ are the NLS (or WNLS) estimates using $y_t$ as the regressand, and let $(\beta^+,\lambda^+)$ be the corresponding estimates when the regressand is $c_0 y_t$ for some $c_0 > 0$. In what follows it is assumed that $x_{t1} \equiv 1$. As shown in the appendix, if the estimates are unique then they must satisfy

$$(6.1) \qquad \lambda^+ = \hat{\lambda}; \quad \beta_1^+ = (c_0^{\hat{\lambda}} - 1)\hat{\lambda}^{-1} + c_0^{\hat{\lambda}}\hat{\beta}_1; \quad \beta_j^+ = c_0^{\hat{\lambda}}\hat{\beta}_j, \ j=2,\ldots,K.$$

The estimate of $\lambda$ is invariant to the rescaling of $y_t$, and the estimates of the other coefficients change so that the fitted values and residuals in the scaled regression are simply scaled up versions of the fitted values and residuals in the unscaled regression. Further, using (6.1) it follows that

$$(6.2) \qquad 1 + \lambda^+ \mathbf{x}\beta^+ = c_0^{\hat{\lambda}}[1 + \hat{\lambda}\mathbf{x}\hat{\beta}];$$

plugging this into (3.14) yields

$$(6.3) \qquad \nabla_\beta \mu_t(\beta^+,\lambda^+) = c_0^{1-\hat{\lambda}} \nabla_\beta \mu_t(\hat{\beta},\hat{\lambda}).$$

Because the residuals are scaled up by $c_0$ and the coefficients are related by (6.1), (6.3) might lead one to believe that t-statistics of the slope coefficients $\beta_2^+,\ldots,\beta_K^+$ are scale invariant. This is indeed true if $\lambda$ has been <u>fixed</u> at an a priori value (e.g. $\lambda = 0$ or $1/2$ or $1$) rather than estimated. However, the gradients with respect to $\lambda$ for the scaled and unscaled models are related by

$$(6.4) \qquad \nabla_\lambda \mu_t(\beta^+,\lambda^+) = c_0 \nabla_\lambda \mu_t(\hat{\beta},\hat{\lambda}) + \hat{\lambda}^{-2} c_0^{1-\hat{\lambda}}[1 + \hat{\lambda}\mathbf{x}_t\hat{\beta}]^{(1/\hat{\lambda})-1}[c_0^{\hat{\lambda}} + \hat{\lambda}\log(c_0)c_0^{\hat{\lambda}}(1+\hat{\lambda}\mathbf{x}_t\hat{\beta}) - 1].$$

Although the second term on the right hand side of (6.4) has zero sample

average (by the first order condition for $(\hat{\beta},\hat{\lambda})$), this term gets squared and then summed in the computation of the standard errors; consequently, the standard errors of $\beta_2^+,\ldots,\beta_K^+$ are not simply scaled down by $c_0^{\hat{\lambda}}$ when $\lambda$ has been estimated along with $\beta$. Interestingly, as shown in the appendix, the lagrange multiplier statistic for exlusion of any 1xQ vector $z_t$ (see model 4.10)) is invariant to the scaling of $y_t$. Consequently, the LM test for the null

$$H_0: \beta_j = 0$$

(j=2,...,K) is scale invariant, whereas the Wald test (based on the t-statistic) is not.

For the standard Box-Cox model, Schesselman (1971) argued that the estimate and standard error of $\lambda$ is scale invariant. This is also the case here; in fact, the robust standard error is also scale invariant. Both of these assertions are proven in the appendix.

One solution to the scale invariance problem for the t-statistics is to add an additional scaling parameter to the conditional mean model. In place of (3.1) consider the model

$$E(y|\mathbf{x}) = \nu[1 + \lambda\mathbf{x}\beta]^{1/\lambda}, \tag{6.5}$$

where $\nu \equiv \exp[E(\log y)]$ is the population geometric mean of y (this requires $P(y > 0) = 1$). Then, scaling y up or down simply alters the scale parameter $\nu$; $\beta$ and $\lambda$ are unchanged. However, model (6.5) cannot be directly estimated by NLS because, if $\mathbf{x}$ contains unity, the parameters $\beta$, $\lambda$, and $\nu$ are not separately identifiable from the NLS objective function. Nevertheless, (6.1) can be easily operationalized by replacing $\nu$ by its sample counterpart $\hat{\nu} \equiv \exp\left[N^{-1}\sum_{t=1}^{N}\log(y_t)\right]$, and $\beta$ and $\lambda$ can be estimated by solving

$$(6.6) \qquad \min_{\beta,\lambda} \sum_{t=1}^{N} (y_t/\hat{\nu} - [1 + \lambda x\beta]^{1/\lambda})^2;$$

each observation on $y_t$ is simply divided by the sample geometric mean of $\{y_t: t=1,\ldots,N\}$, and then the model in section 3 is estimated. The estimates of $\beta$ are trivially scale invariant because $y_t/\hat{\nu}$ is invariant to scaling. Spitzer (1984) recommends the same strategy for the Box-Cox model. There is, however, a somewhat subtle issue that needs to be addressed in implementing this procedure. The solutions $\hat{\beta}$ and $\hat{\lambda}$ to (6.6) depend on the estimator $\hat{\nu}$. Although it is tempting to ignore the randomness of $\hat{\nu}$, the estimator of the asymptotic variance of $\hat{\theta} \equiv (\hat{\beta},\hat{\lambda})$ should reflect this additional source of uncertainty (as different samples of $\{y_t\}$ are obtained, the estimator $\hat{\nu}$ generally changes). In the general WNLS case, the easiest approach to this problem is to view $\hat{\theta}$ is a "two-step" estimator that solves

$$(6.7) \qquad \min_{\theta} \sum_{t=1}^{N} (y_t - \mu(x_t;\theta,\hat{\nu}))^2/\hat{\omega}_t$$

where

$$\mu(x,\theta,\nu) \equiv \nu[1 + \lambda x\beta]^{1/\lambda}.$$

The appendix derives the asymptotic variance of the solution $\hat{\theta}$ of (6.7) which accounts for the variability of $\hat{\nu}$. Define the (K+1)x1 vector

$$\tilde{c}_N \equiv N^{-1} \sum_{t=1}^{N} (\nabla_\theta \hat{\mu}_t/\sqrt{\hat{\omega}_t})' (\nabla_\nu \hat{\mu}_t/\sqrt{\hat{\omega}_t}),$$

where $\nabla_\theta \hat{\mu}_t$ is the same as derived in section 3 <u>except</u> that it is now multiplied by $\hat{\nu}$; also, note that $\nabla_\nu \hat{\mu}_t = [1 + \hat{\lambda} x \hat{\beta}]^{1/\hat{\lambda}}$ is simply the fitted value for the <u>scaled</u> regressand $y_t$. A consistent estimate of the asymptotic variance of $\hat{\theta}$ is

$$(6.8) \qquad \left(\sum_{t=1}^{N} \nabla_\theta \tilde{\mu}_t' \nabla_\theta \tilde{\mu}_t\right)^{-1} [I_P | -\tilde{c}_N] \left(\sum_{t=1}^{N} \tilde{s}_t' \tilde{s}_t\right) [I_P | -\tilde{c}_N]' \left(\sum_{t=1}^{N} \nabla_\theta \tilde{\mu}_t' \nabla_\theta \tilde{\mu}_t\right)^{-1},$$

where $P \equiv K+1$, $\tilde{s}_t$ is the $1x(P+1)$ vector

$$\tilde{s}_t \equiv (\tilde{\epsilon}_t \nabla_\theta \tilde{\mu}_t, \hat{\nu}\log(y_t/\hat{\nu})),$$

$\hat{\epsilon}_t \equiv y_t - \mu_t(\hat{\theta},\hat{\nu}) \equiv y_t - \hat{\nu}[1 + \hat{\lambda}x_t\hat{\beta}]^{1/\lambda}$, $\nabla_\theta \tilde{\mu}_t \equiv \nabla_\theta \hat{\mu}_t/\sqrt{\hat{\omega}_t}$, and $\tilde{\epsilon}_t \equiv \hat{\epsilon}_t/\sqrt{\hat{\omega}_t}$. The estimator (6.8) is also robust to variance misspecification (heteroskedasticity when $\hat{\omega}_t \equiv 1$). A degrees of freedom adjustment would scale (6.8) up by the factor N/(N-P). Note that in the construction of $\tilde{s}_t$, $\hat{\nu}\log(y_t/\hat{\nu})$ is <u>not</u> weighted by $1/\sqrt{\hat{\omega}_t}$.

Generally speaking, (6.8) differs from the usual robust covariance matrix estimator

$$(6.9) \qquad \left(\sum_{t=1}^{N} \nabla_\theta \tilde{\mu}_t' \nabla_\theta \tilde{\mu}_t\right)^{-1} \left(\sum_{t=1}^{N} \tilde{\epsilon}_t^2 \nabla_\theta \tilde{\mu}_t' \nabla_\theta \tilde{\mu}_t\right) \left(\sum_{t=1}^{N} \nabla_\theta \tilde{\mu}_t' \nabla_\theta \tilde{\mu}_t\right)^{-1}.$$

however, as shown in the appendix, (6.8) and (6.9) produce numerically identical estimates for $se(\hat{\lambda})$; this is as it should be because the theoretical asymptotic variance of $\hat{\lambda}$ is unaffected by the estimation of $\nu$. The scale corrected standard errors of $\hat{\beta}_1,\ldots,\hat{\beta}_K$ from (6.8) will generally be different from those obtained from (6.9), reflecting the influence of the variation in the estimator $\hat{\nu}$.

Similar conclusions can be obtained for the general model (5.1) and (5.2). As in section 5, let $x_{tj}$, $j=J+1,\ldots,K$ denote strictly positive regressors that Box-Cox transformations are to be applied to. Then the t-statistics of $\hat{\beta}_{J+1},\ldots,\hat{\beta}_K$ are <u>not</u> invariant to the scaling of $x_{J+1},\ldots,x_K$, even if y is not scaled. The estimates of $\beta_2,\ldots,\beta_J$ are invariant, as are the associated t-statistics; $\hat{\beta}_1$ and its t-statistic are essentially never invariant. It can be shown that if $\rho_j$ is fixed at a particular value, rather than estimated, then $t(\hat{\beta}_j)$ is invariant to scaling of $x_j$. The estimates of $\lambda,\rho_{J+1},\ldots,\rho_K$ are invariant to scaling y and/or $x_{J+1},\ldots,x_K$, as is the

asyptotic variance matrix of the subvector $(\hat{\lambda},\hat{\rho}_{J+1},\ldots,\hat{\rho}_K)'$. Again, the appendix verifies these claims.

Lagrange multiplier tests for any exclusion restrictions <u>are</u> again scale invariant, so these can be used as alternatives to testing the exclusion of particular variables via t-statistics. The appendix contains a proof of this assertion.

The scaled version of (5.1) becomes

$$\mu(\mathbf{x};\theta,\nu,\eta) \equiv \nu[1 + \lambda\mathbf{x}(\rho,\eta)\beta]^{1/\lambda}, \quad \lambda \neq 0 \tag{6.10}$$

$$\equiv \exp[\mathbf{x}(\rho,\eta)\beta], \quad \lambda = 0 \tag{6.11}$$

where $\mathbf{x}(\rho,\eta) \equiv (x_1,x_2,\ldots,x_J,x_{J+1}(\rho_{J+1},\eta_{J+1}),\ldots,x_{J+1}(\rho_{J+1},\eta_{J+1}))$, $x_j(\rho_j,\eta_j)$ denotes the Box-Cox transformation of the scaled variable $x_j/\eta_j$, $\eta_j \equiv \exp[E(\log x_j)]$ is the population geometric mean of $x_j$, j=J+1,...,.K, and $\theta \equiv (\beta',\lambda,\rho')'$ is now a K+1+(K-J)x1 vector (define P $\equiv$ K+1+(K-J)). The first stage estimators are now $\hat{\nu}$, the sample geometric mean of $\{y_t\}$, and $\hat{\eta}_j$, j=J+1,...,K, the sample geometric means of $\{x_{tj}\}$, j=J+1,...,K. The WNLS estimator now solves

$$\min_{\theta} \sum_{t=1}^{N} (y_t - \mu(\mathbf{x}_t;\theta,\hat{\nu},\hat{\eta}))^2, \tag{6.12}$$

which is algebraically the same as first scaling y and $x_{J+1},\ldots,x_K$ by their sample geometric means and estimating the regression function as in section 5. Collecting the "nuisance" parameters into the 1+(K-J)x1 vector $\pi \equiv (\nu,\eta')'$, the gradient $\nabla_\pi \mu_t(\theta,\pi)$ is needed to compute the correct asymptotic covariance matrix. But (for $\lambda \neq 0$),

$$\nabla_{\nu}\mu_t(\theta,\pi) = [1 + \lambda x_t(\rho,\eta)\beta]^{1/\lambda}, \tag{6.13}$$

$$\nabla_{\eta_j}\mu_t(\theta,\pi) = -\nu[1 + \lambda x_t(\rho,\eta)\beta]^{\{(1/\lambda)-1\}}(\delta_j/\eta_j)(x_{tj}/\eta_j)^{\rho_j}, \quad j=J+1,\ldots,K. \tag{6.14}$$

Let $\nabla_{\pi}\mu_t(\theta,\pi)$ denote the 1x(K-J)+1 row vector consisting of these elements. Let $\tilde{C}_N$ now denote the Px(1+K-J) matrix

$$\tilde{C}_N \equiv N^{-1}\sum_{t=1}^{N}(\nabla_{\theta}\hat{\mu}_t/\sqrt{\hat{\omega}_t})'(\nabla_{\pi}\hat{\mu}_t/\sqrt{\hat{\omega}_t}), \tag{6.15}$$

and let $\tilde{s}_t$ denote the 1x[P+(1+K-J)] vector

$$\tilde{s}_t \equiv (\tilde{\epsilon}_t\nabla_{\theta}\tilde{\mu}_t, \hat{\nu}\log(y_t/\hat{\nu}), \hat{\eta}_{J+1}\log(x_{t,J+1}/\hat{\eta}_{J+1}), \ldots, \hat{\eta}_K\log(x_{tK}/\hat{\eta}_{J+1})).$$

As usual, $\hat{\epsilon}_t \equiv y_t - \mu_t(\hat{\theta},\hat{\nu},\hat{\eta}) \equiv y_t - \hat{\nu}[1 + \hat{\lambda}x_t(\hat{\rho},\hat{\eta})\hat{\beta}]^{1/\hat{\lambda}}$, $\nabla_{\theta}\tilde{\mu}_t \equiv \nabla_{\theta}\hat{\mu}_t/\sqrt{\hat{\omega}_t}$, and $\tilde{\epsilon}_t \equiv \hat{\epsilon}_t/\sqrt{\hat{\omega}_t}$. Then a consistent estimator of the asymptotic variance of the WNLS estimator $\hat{\theta}$ is

$$\left(\sum_{t=1}^{N}\nabla_{\theta}\tilde{\mu}_t'\nabla_{\theta}\tilde{\mu}_t\right)^{-1}[I_P|-\tilde{C}_N]\left(\sum_{t=1}^{N}\tilde{s}_t'\tilde{s}_t\right)[I_P|-\tilde{C}_N]'\left(\sum_{t=1}^{N}\nabla_{\theta}\tilde{\mu}_t'\nabla_{\theta}\tilde{\mu}_t\right)^{-1}, \tag{6.16}$$

which is the same form as (6.8) once $\nabla_{\theta}\tilde{\mu}_t$, $\tilde{s}_t$, and $\tilde{C}_N$ have been appropriately modified. As with the simple model (3.1), scaling $y_t$ and/or $x_{tj}$, $j=J+1,\ldots,K$ does not affect $\hat{\lambda}$, $\hat{\rho}_{J+1},\ldots,\hat{\rho}_K$ or their standard errors; this is reflected in the fact that the variance-covariance matrix estimator of the subvector $(\hat{\lambda},\hat{\rho}_{J+1},\ldots,\hat{\rho}_K)$ obtained from (6.16) is identical to the robust formula which ignores the correction factor (see the appendix for a proof). Again, the scale corrected t-statistics of $\hat{\beta}_1,\ldots,\hat{\beta}_K$ will generally be different from the robust t-statistics which ignore the estimation of the scale parameters.

The formula for the asymptotic covariance matrix estimator when $\lambda$ is restricted to be zero (or any other fixed value) is obtained by omitting $\nu$

from the model and redefining the nuisance parameters to be $\pi \equiv \eta$. The gradients $\nabla_\theta \mu_t(\theta,\pi)$ and $\nabla_\pi \mu_t(\theta,\pi)$ are also appropriately redefined. In this case -- at least for cross section or static time series applications -- one might treat $\hat{\eta}$ as nonrandom if the $x_t$ are being treated as nonrandom. The usual robust formula would be asymptotically valid in this case.

## 7. Some Practical Considerations

The model presented in section 5 contains as special cases many of the functional forms used by applied econometricians in studies involving nonnegative variables. The generality obtained means that more work is involved in selecting an acceptable model. In addition, the problem of choosing the weights to compute the WNLS estimators is important for actual implementation. Although every application has unique features, some general guidelines can be given.

First, one has to decide on which restricted version of the conditional mean function (5.1) to start with. For computational reasons this would almost always involve $\lambda = 1$ or $\lambda = 0$, possibly along with other constraints on the $\rho_j$. Because the constant elasticity and constant semi-elasticity forms are so appealing for nonnegative variables, a good starting point is $\lambda = 0$ and $\rho_j = 0$, $j=J+1,\ldots,K$, or $\lambda = 0$ and $\rho_j = 1$, $j=J+1,\ldots,K$. Then $\beta$ can be initially estimated by NLS of an exponential regression function, which is relatively easy.

If $y_t$ is continuously distributed on $[0,\infty)$ then one sensible choice of $\hat{\omega}_t$ is the the square of the fitted values from the NLS estimation. This is the optimal choice if y given x has an exponential distribution or if log y given x is normal with constant variance. As these distributional

assumptions are leading cases for nonnegative continuously distributed variables, this choice of weights makes some sense. (But recall that the analysis in sections 3, 4, and 5 does not actually require that the weighting function be proportional to $V(y|\mathbf{x})$.) If y is a count variable then a sensible choice of weights is simply the predicted values from the NLS estimation, as this is optimal when y given $\mathbf{x}$ has a Poisson distribution. Rather than performing the WNLS estimation in two steps, quasi-maximum likelihood estimation using the Exponential or Poisson distributions can be implemented directly. A more flexible set of weights can be obtained as the fitted values from the OLS regression

$$\hat{\epsilon}_t^2 \text{ on } 1,\ \mu(\mathbf{x}_t,\hat{\theta}),\ [\mu(\mathbf{x}_t,\hat{\theta})]^2, \tag{7.1}$$

where the fitted values $\mu(\mathbf{x}_t,\hat{\theta})$ and residuals $\hat{\epsilon}_t$ would most likely come from an initial NLS estimation. This contains as a special case the optimal weights for a geometric distribution (which is in the linear exponential family), where, asymptotically, the intercept in regression (7.1) would be zero and the coefficients on $\mu(\mathbf{x}_t,\hat{\theta})$ and $[\mu(\mathbf{x}_t,\hat{\theta})]^2$ would both be unity.

There is no reason to restrict the weights to functions of the fitted values; a simple and fairly flexible approach is to exponentiate the fitted values from the regression

$$\log \hat{\epsilon}_t^2 \text{ on } 1,\ x_{t2},\ \ldots,\ x_{tK},\ t=1,\ldots,N,$$

where $\hat{\epsilon}_t$ are the NLS residuals. While this choice of $\hat{\omega}_t$ need not produce consistent estimates of $V(y|\mathbf{x})$ (up to scale), it could still improve the precision of the WNLS estimator relative to that of the NLS estimator.

It should be emphasized that weighted NLS need only be considered as an alternative to NLS if the researcher feels that the variance of the NLS

estimator is too large. As a test of model specification, however, it is frequently useful to compute the NLS and a weighted NLS estimator and compare the two via a Hausman test. Wooldridge (1990) covers robust, regression-based Hausman tests that apply in this context.

Once a set of weights has been selected, the LM tests developed in section 5 can be used to test $H_0$: $\lambda = 0$, $\rho_j = 0$, j=J+1,...,K or $H_0$: $\lambda = 0$, $\rho_j = 1$, j=J+1,...,K. If both of these hypotheses are rejected then the less restrictive hypothesis $H_0$: $\lambda = 0$ can be tested. If model (5.2) is rejected entirely then one might turn to the model with $\lambda = 1$. The hypotheses $H_0$: $\lambda = 1$, $\rho_j = 0$, j=J+1,...,K, $H_0$: $\lambda = 1$, $\rho_j = 1$, j=J+1,...,K, and $H_0$: $\lambda = 1$ are of particular interest. All of these LM tests are invariant to the scaling of y and/or **x**.

If all versions of the models $\lambda = 0$ and $\lambda = 1$ are rejected then perhaps the unrestricted model needs to be estimated. The issue of lack of scale invariance of the t-statistics then becomes an issue, and the methods of section 6 can be used. Of course nothing guarantees that the general model (5.1) is correctly specified for $E(y|\mathbf{x})$; the Hausman test can be applied to test for misspecification of the general model.

Another problem arises if the data cannot reject either $\lambda = 0$ or $\lambda = 1$. In this case one might turn to a nonnested hypotheses test to attempt to distinguish between the two models. Computationally simple robust nonnested hypotheses tests are discussed in Wooldridge (1990). A simple goodness of fit criterion is to choose the model with the smallest SSR from NLS estimation or from WNLS estimation using the same set of weights.

## 8. Concluding Remarks

The Box-Cox transformation has proven to be a useful tool for generalizing functional form in statistics and econometrics. It is not, however, well-suited for applications where interest centers on $E(y|x)$ rather than on the conditional expectation of some nonlinear transformation of $y$. When $y$ is the quantity of interest to economic agents and policy makers it is important to have available estimates of $E(y|x)$ that are easy to compute and robust to distributional misspecification. Estimating a linear model where the regressand is, say, $y^{1/2}$, is not very useful unless $E(y|x)$ can be recovered from $E(y^{1/2}|x)$. In the Box-Cox framework with $\lambda = 1/2$ computation of $E(y|x)$ requires normality and homoskedasticity of $y^{1/2}$.

Requiring that some power transformation simultaneously induce linearity of the conditional expectation, homoskedasticity, and normality is asking a lot of economic data, and is not in itself important for estimating economic quantities. This paper has proposed as an alternative estimating a nonlinear model for $E(y|x)$ that is flexible enough to contain several special cases that are used frequently by applied researchers. Further, no second moment or other distributional assumptions are relied upon to obtain consistent estimates or to perform asymptotically valid inference. As a consequence all of the robust LM tests of the special cases covered in sections 3 and 5 are pure conditional mean tests: a rejection can be confidently interpreted as a rejection of the model for $E(y|x)$, and not as a rejection of some other less important distributional assumption.

References

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

Andrews, D.F. (1971), "A Note on the Selection of Data Transformations," *Biometrika*, 58, 249-254.

Berndt, E.R. and M.S. Khaled (1979), "Parametric Productivity Measurement and Choice among Flexible Functional Forms," *Journal of Political Economy*, 87, 1220-1245.

Bickel, P.J. and K.A. Doksum (1981), "An Analysis of Transformations Revisited," *Journal of the American Statistical Association*, 76, 296-311.

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

Cohen, A. and H.B. Sackrowitz (1987), "An Approach to Inference Following Model Selection With Applications to Transformation-Based and Adaptive Inference," *Journal of the American Statistical Association*, 82, 1123-1130.

Davidson, R. and J.G. MacKinnon (1985a), "Testing Linear and Loglinear Regressions Against Box-Cox Alternatives," *Canadian Journal of Economics*, 25, 499-517.

Davidson, R. and J.G. MacKinnon (1985b), "Heteroskedasticity-robust Tests in Regression Directions," *Annales de l'INSEE*, 59/60, 183-218.

Goldberger, A.S. (1968), "The Interpretation and Estimation of Cobb-Douglas Production Functions," *Econometrica*, 35, 464-472.

Gourieroux, C., A. Monfort, and A. Trognon (1984), "Pseudo Maximum Likelihood Methods: Theory," *Econometrica*, 52, 681-700.

Hausman, J.A., Hall, B.H., and Z. Griliches (1984), "Econometric Models for Count Data with an Application to the Patents-R&D Relationship," *Econometrica*, 52,

Hinkley D.V. and G. Runger (1984), "The Analysis of Transformed Data," *Journal of the American Statistical Association*, 79, 302-309.

Huang, C.J and O.R. Grawe (1980), "Functional Forms and the Demand for Meat in the U.S.: A Comment," *Review of Economics and Statistics*, 62, 144-146.

Huang, C.J and J.A. Kelingos (1979), "Conditional Mean Function and a General Specification of the Disturbance in Regression," *Southern Economic Journal*, 45, 710-717.

Judge, G.G., Griffiths, W.E., Hill, R.C., Lütkepohl, H., and Lee, T.-C. (1985), *The Theory and Practice of Econometrics*. New York: Wiley, Second Edition.

MacKinnon, J.G. and L. McGee (1989), "Transforming the Dependent Variable in Regression Models," forthcoming, *International Economic Review*.

Mizon, G.E. (1977), "Inferential Procedures in Nonlinear Models: An Application in a UK Industrial Cross Section Study of Factor Substitution and Returns to Scale," *Econometrica*, 45, 1221-1242.

Mukerji, V. (1963), "A Generalized SMAC Function with Constant Ratios of Elasticity of Substitution," *Review of Economic Studies*, 30, 233-236.

Nelson, H.L. and C.W.J. Granger (1979), "Experience with Using the Box-Cox Transformation for Forecasting Economic Time Series," *Journal of Eonometrics*, 10, 57-69.

Newey, W.K. and K. West (1987), "A Simple Positive Semi-Definite Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703-708.

Papke, L.E. (1989), "The Influence of Interstate Tax Differentials on the Birth of New Firms: Estimates of a Poisson Process," Boston University School of Management Working Paper 89-01.

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

Poirier, D.J. and A. Melino (1978), "A Note on the Interpretation of Regression Coefficients Within a Class of Truncated Distributions," *Econometrica*, 46, 1207-1209.

Poirier, D.J. and P.A. Ruud (1979), "A Simple Lagrange Multiplier Test for Lognormal Regression," *Economics Letters*, 4, 251-255.

Poirier, D.J. and P.A. Ruud (1983), "Diagnostic Testing in Missing Data Models," *International Economic Review*, 24, 537-546.

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

Seaks, T.G. and S.K. Layson (1983), "Box-Cox Estimation with Standard Econometric Problems," *Review of Economics and Statistics*, 65, 160-164.

Spitzer, J.J. (1982), "A Primer on Box-Cox Estimation," *Review of Economics and Statistics*, 64, 307-313.

Spitzer, J.J. (1984), "Variance Estimates in Models with Box-Cox Transformations: Implications for Estimation and Hypothesis Testing," *Review of Economics and Statistics*, 66, 645-652.

White, H. (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817-838.

White, H. and I. Domowitz (1984), "Nonlinear Regression with Dependent Observations," *Econometrica*, 52, 143-162.

Wooldridge, J.M. (1989), "Specification Testing and Quasi-Maximum Likelihood Estimation," MIT Department of Economics Working Paper 479.

Wooldridge, J.M. (1990), "On the Application of Robust, Regression-Based Diagnostics to Models of Conditional Means and Conditional Variances," mimeo, MIT Department of Economics. Forthcoming, *Journal of Econometrics* special issue on "Diagnostic Testing."

The results in this appendix require that the vector of explanatory variables contains a constant. Thus, $x_{t1} \equiv 1$, $t=1,2,\ldots$ is assumed throughout. For notational simplicity, the results are proven for the unweighted case; it is obvious how the proofs are modified for weighted NLS. Claims 1 - 4 pertain to the model of section 3.

Claim 1: $\lambda^+ = \hat{\lambda}$; $\beta_1^+ = (c_0^{\hat{\lambda}} - 1)\hat{\lambda}^{-1} + c_0^{\hat{\lambda}}\hat{\beta}_1$; $\beta_j^+ = c_0^{\hat{\lambda}}\hat{\beta}_j$, $j=2,\ldots,K$.

Proof: Let $\mu_t(\beta,\lambda) \equiv [1 + \lambda x_t\beta]^{1/\lambda}$, and let $\nabla_\beta\mu(\beta,\lambda)$ and $\nabla_\lambda\mu(\beta,\lambda)$ denote the derivatives. Then the first order conditions for $(\beta^+,\lambda^+)$ are

$$\sum_{t=1}^{N} \nabla_\beta\mu(\beta^+,\lambda^+)'(c_0y_t - \mu_t(\beta^+,\lambda^+)) \equiv 0, \qquad (a.1)$$

$$\sum_{t=1}^{N} \nabla_\lambda\mu(\beta^+,\lambda^+)'(c_0y_t - \mu_t(\beta^+,\lambda^+)) \equiv 0. \qquad (a.2)$$

Because the solutions are assumed to be unique, it suffices to show that $\beta^+$ and $\lambda^+$ given by (6.1) satisfy (a.1) and (a.2). Then (a.1) reduces to showing

$$\sum_{t=1}^{N} c_0^{1-\hat{\lambda}}\nabla_\beta\mu(\hat{\beta},\hat{\lambda})'(c_0y_t - c_0\mu_t(\hat{\beta},\hat{\lambda}))$$

$$= c_0^{2-\hat{\lambda}} \sum_{t=1}^{N} \nabla_\beta\mu(\hat{\beta},\hat{\lambda})'(y_t - \mu_t(\hat{\beta},\hat{\lambda})) = 0. \qquad (a.3)$$

But (a.3) follows from the first order condition for $(\hat{\beta},\hat{\lambda})$. Next, from (6.4),

$$\sum_{t=1}^{N} \nabla_\lambda\mu(\beta^+,\lambda^+)'(c_0y_t - \mu_t(\beta^+,\lambda^+)) = \sum_{t=1}^{N} \{c_0\nabla_\lambda\mu_t(\hat{\beta},\hat{\lambda}) +$$

$$\hat{\lambda}^{-2}c_0^{1-\hat{\lambda}}[1+\hat{\lambda}x_t\hat{\beta}]^{(1/\hat{\lambda})-1}[c_0^{\hat{\lambda}}+\hat{\lambda}\log(c_0)c_0^{\hat{\lambda}}(1+\hat{\lambda}x_t\hat{\beta})-1]\}(c_0y_t-c_0\mu_t(\hat{\beta},\hat{\lambda}))$$

$$= c_0^2 \sum_{t=1}^{N} \nabla_\lambda \mu_t(\hat{\beta},\hat{\lambda})(y_t - \mu_t(\hat{\beta},\hat{\lambda})) \tag{a.4}$$

$$+ \sum_{t=1}^{N} \hat{\lambda}^{-2} c_0^{2-\hat{\lambda}} [1+\hat{\lambda}x_t\hat{\beta}]^{(1/\hat{\lambda})-1} [c_0^{\hat{\lambda}} + \hat{\lambda}\log(c_0)c_0^{\hat{\lambda}}(1+\hat{\lambda}x_t\hat{\beta})-1]\}(y_t - \mu_t(\hat{\beta},\hat{\lambda})).$$

The first term in (a.4) is zero by the first order condition for $(\hat{\beta},\hat{\lambda})$. The first order condition for $(\hat{\beta},\hat{\lambda})$ also implies that the second part of (a.4) is zero. This is because

$$\sum_{t=1}^{N} [1 + \hat{\lambda}x_t\hat{\beta}]^{\{(1/\hat{\lambda})-1\}}(y_t - \mu_t(\hat{\beta},\hat{\lambda})) \equiv 0 \tag{a.5}$$

is the first element of $\sum_{t=1}^{N} \nabla_\beta \mu_t(\hat{\beta},\hat{\lambda})'(y_t - \mu_t(\hat{\beta},\hat{\lambda}))$ if $x_{t1} \equiv 1$. Also,

$$\sum_{t=1}^{N} [1 + \hat{\lambda}x_t\hat{\beta}]^{\{(1/\hat{\lambda})-1\}}(1 + \hat{\lambda}x_t\hat{\beta})(y_t - \mu_t(\hat{\beta},\hat{\lambda}))$$

$$= \sum_{t=1}^{N} [1 + \hat{\lambda}x_t\hat{\beta}]^{\{(1/\hat{\lambda})-1\}}(y_t - \mu_t(\hat{\beta},\hat{\lambda}))$$

$$+ \hat{\lambda} \sum_{t=1}^{N} [1 + \hat{\lambda}x_t\hat{\beta}]^{\{(1/\hat{\lambda})-1\}}x_t\hat{\beta}(y_t - \mu_t(\hat{\beta},\hat{\lambda})) = 0$$

since this is a linear combination of $\sum_{t=1}^{N} \nabla_\beta \mu_t(\hat{\beta},\hat{\lambda})'(y_t - \mu_t(\hat{\beta},\hat{\lambda}))$. This establishes (a.1) and (a.2) for $(\beta^+,\lambda^+)$ given by (6.4), and completes the proof. ■

Claim 2: $se(\hat{\lambda})$ is scale invariant.

Proof: It is shown that the asymptotic variance of $\hat{\lambda}$ is scale invariant. A standard mean-value expansion yields

$$\sqrt{N}(\hat{\theta} - \theta) = \left[N^{-1} \sum_{t=1}^{N} \nabla_\theta \mu_t(\theta)' \nabla_\theta \mu_t(\theta)\right]^{-1} N^{-1/2} \sum_{t=1}^{N} \nabla_\theta \mu_t(\theta)' \epsilon_t + o_p(1). \tag{a.6}$$

Focusing on the last element of (a.6), that corresponding to $\lambda$, we have

$$\sqrt{N}(\hat{\lambda} - \lambda) = \left(N^{-1}\sum_{t=1}^{N} r_t^2\right)^{-1} N^{-1/2}\sum_{t=1}^{N} r_t \epsilon_t + o_p(1), \qquad (a.7)$$

where $r_t$ is the residual from the regression

$$\nabla_\lambda \mu_t(\theta) \quad \text{on} \quad \nabla_\beta \mu_t(\theta), \quad t=1,\ldots,N. \qquad (a.8)$$

The true error $\epsilon_t$ is scaled up by $c_0$ when $y_t$ is scaled up by $c_0$, so it suffices to show that $r_t$ is also scaled up by $c_0$. Then the expression (a.7) is scale invariant, and then so must be the asymptotic variance of $\hat{\lambda}$. Let values superscripted by "o" denote the scaled values. Then

$$\nabla_\beta \mu_t(\theta^o) = c_0^{1-\lambda} \nabla_\beta \mu_t(\theta) \qquad (a.9)$$

$$\nabla_\lambda \mu_t(\theta^o) = c_0 \nabla_\lambda \mu_t(\theta) +$$

$$\lambda^{-2} c_0^{1-\lambda} [1 + \lambda x_t \beta]^{(1/\lambda)-1} [c_0^\lambda + \lambda \log(c_0) c_0^\lambda (1+\lambda x_t \beta) - 1]. \qquad (a.10)$$

From (a.10), $\nabla_\lambda \mu_t(\theta^o)$ can be expressed succinctly as $\nabla_\lambda \mu_t(\theta^o) = c_0 \nabla_\lambda \mu_t(\theta) + \nabla_\beta \mu_t(\theta) a$ for a Kxl vector a. From (a.9),

$$\nabla_\lambda \mu_t(\theta^o) = c_0 \nabla_\lambda \mu_t(\theta) + c_0^{\lambda-1} \nabla_\beta \mu_t(\theta^o) a.$$

Consequently, the residuals from the regression

$$\nabla_\lambda \mu_t(\theta^o) \quad \text{on} \quad \nabla_\beta \mu_t(\theta^o),$$

say $r_t^o$, satisfy $r_t^o = c_0 r_t$; thus, $\left(N^{-1}\sum_{t=1}^{N} (r_t^o)^2\right)^{-1} N^{-1/2} \sum_{t=1}^{N} r_t^o \epsilon_t^o = \left(N^{-1}\sum_{t=1}^{N} r_t^2\right)^{-1} N^{-1/2} \sum_{t=1}^{N} r_t \epsilon_t$, and the asymptotic variance of $\hat{\lambda}$ is invariant. That the computed standard errors are invariant follows because $\hat{r}_t$ and $\hat{\epsilon}_t$ are also scaled up by $c_0$. ■

Claim 3: The LM statistic for exclusion restrictions is scale invariant.

Proof: Consider the unrestricted model

$$\mu(x,z;\beta,\lambda,\delta) = [1 + \lambda x\beta + \lambda z\delta]^{1/\lambda},$$

and consider testing $H_0$: $\delta = 0$. If the regressand is $c_0 y_t$ for $c_0 > 0$ then the

gradients used for the test on scaled data are related by

$$\nabla_{\beta}\mu_t(\beta^+,\lambda^+,0) = c_0^{1-\hat{\lambda}}\nabla_{\beta}\mu_t(\hat{\beta},\hat{\lambda},0)$$

$$\nabla_{\lambda}\mu_t(\beta^+,\lambda^+,0) = c_0\nabla_{\lambda}\mu_t(\hat{\beta},\hat{\lambda},0) +$$

$$\hat{\lambda}^{-2}c_0^{1-\hat{\lambda}}[1 + \hat{\lambda}x_t\hat{\beta}]^{(1/\hat{\lambda})-1}[c_0^{\hat{\lambda}} + \hat{\lambda}\log(c_0)c_0^{\hat{\lambda}}(1+\hat{\lambda}x_t\hat{\beta}) - 1]$$

$$\equiv c_0\nabla_{\lambda}\mu_t(\hat{\beta},\hat{\lambda},0) + \nabla_{\beta}\mu_t(\hat{\beta},\hat{\lambda},0)\hat{a}$$

$$\nabla_{\delta}\mu_t(\beta^+,\lambda^+,0) = c_0^{1-\hat{\lambda}}\nabla_{\alpha}\mu_t(\hat{\beta},\hat{\lambda},0).$$

Because the gradients of the scaled data are linear combinations of the gradients for the unscaled data, and because $\epsilon_t^+ = c_0\hat{\epsilon}_t$, the r-squareds from the regressions

$$\epsilon_t^+ \text{ on } \nabla_{\beta}\mu_t(\beta^+,\lambda^+,0),\ \nabla_{\lambda}\mu_t(\beta^+,\lambda^+,0),\ \nabla_{\delta}\mu_t(\beta^+,\lambda^+,0)$$

and

$$\hat{\epsilon}_t \text{ on } \nabla_{\beta}\mu_t(\hat{\beta},\hat{\lambda},0),\ \nabla_{\lambda}\mu_t(\hat{\beta},\hat{\lambda},0),\ \nabla_{\delta}\mu_t(\hat{\beta},\hat{\lambda},0)$$

are numerically identical. This shows that the nonrobust LM statistics are numerically identical. For the robust test, note that the residuals $r_t^+$ from the regression

$$\nabla_{\delta}\mu_t(\beta^+,\lambda^+,0) \text{ on } \nabla_{\beta}\mu_t(\beta^+,\lambda^+,0),\ \nabla_{\lambda}\mu_t(\beta^+,\lambda^+,0)$$

are scalar multiples of the residuals $\hat{r}_t$ from the regression

$$\nabla_{\delta}\mu_t(\hat{\beta},\hat{\lambda},0) \text{ on } \nabla_{\beta}\mu_t(\hat{\beta},\hat{\lambda},0),\ \nabla_{\lambda}\mu_t(\hat{\beta},\hat{\lambda},0)$$

($r_t^+ = c_0^{1-\hat{\lambda}}\hat{r}_t$). Consequently, the r-squareds from the regressions

$$1 \text{ on } \epsilon_t^+ r_t^+$$

and

$$1 \text{ on } \hat{\epsilon}_t\hat{r}_t$$

are identical. ■

Claim 4: The scale corrected standard error of $\hat{\lambda}$ is identical to the robust standard error of $\hat{\lambda}$ ($y_t$ has been scaled by the geometric mean $\hat{\nu}$).

Proof: $\hat{\theta} \equiv (\hat{\beta},\hat{\lambda})$ now satisfies

$$\sum_{t=1}^{N} \nabla_\theta \mu_t(\hat{\theta},\hat{\nu})'(y_t - \mu_t(\hat{\theta},\hat{\nu})) \equiv 0$$

where $\mu_t(\hat{\theta},\hat{\nu}) = \hat{\nu}[1 + \hat{\lambda}\mathbf{x}_t\hat{\beta}]^{1/\hat{\lambda}}$ and $\nabla_\beta \mu_t(\hat{\theta},\hat{\nu})$ and $\nabla_\lambda \mu_t(\hat{\theta},\hat{\nu})$ are also scaled up by $\hat{\nu}$. A mean value expansion along with the delta method yields

$$\sqrt{N}(\hat{\theta} - \theta) = \left(N^{-1}\sum_{t=1}^{N} \nabla_\theta \mu_t' \nabla_\theta \mu_t\right)^{-1} N^{-1/2} \sum_{t=1}^{N} \nabla_\theta \mu_t' \epsilon_t \qquad \text{(a.11)}$$

$$- \left(N^{-1}\sum_{t=1}^{N} \nabla_\theta \mu_t' \nabla_\theta \mu_t\right)^{-1}\left(N^{-1}\sum_{t=1}^{N} \nabla_\theta \mu_t' \nabla_\nu \mu_t\right) N^{-1/2} \sum_{t=1}^{N} \nu \log(y_t/\nu) + o_p(1)$$

where all elements are evaluated at $(\theta,\nu)$. The second term in (a.11) is the contribution due to the estimation of $\nu$; the first term is as before. Thus, it suffices to show that last element of

$$\left(N^{-1}\sum_{t=1}^{N} \nabla_\theta \mu_t' \nabla_\theta \mu_t\right)^{-1}\left(N^{-1}\sum_{t=1}^{N} \nabla_\theta \mu_t' \nabla_\nu \mu_t\right) \qquad \text{(a.12)}$$

(the element corresponding to $\lambda$) is identically zero. But (a.12) is the vector of coefficients from the regression

$$\nabla_\nu \mu_t \quad \text{on} \quad \nabla_\theta \mu_t, \quad t=1,\ldots,N$$

or

$$\nabla_\nu \mu_t \quad \text{on} \quad \nabla_\beta \mu_t, \ \nabla_\lambda \mu_t, \quad t=1,\ldots,N.$$

The coefficient on $\nabla_\lambda \mu_t$ is also obtained by first obtaining the residuals $r_t$ from

$$\nabla_\lambda \mu_t \quad \text{on} \quad \nabla_\beta \mu_t, \quad t=1,\ldots,N,$$

and then computing the coefficient from the simple regression

$$\nabla_\nu \mu_t \quad \text{on} \quad r_t.$$

Thus, it suffices to show that $\nabla_\nu \mu_t$ and $r_t$ are orthogonal. But the residuals $r_t$ are orthogonal to any linear combination of $\nabla_\beta \mu_t$, i.e.

$$\sum_{t=1}^{N} \nabla_\beta \mu_t' r_t = 0,$$

and $\nabla_\nu \mu_t = [1 + \lambda x_t \beta]^{1/\lambda} = [1 + \lambda x_t \beta]^{1/\lambda - 1} + [1 + \lambda x_t \beta]^{1/\lambda - 1} x_t \beta$, which is a linear combination of $\nabla_\beta \mu_t$ whenever $x_t$ contains a constant. Thus, for $\hat{\lambda}$, (a.11) is the same as (a.6). This completes the proof. ■

Claims 5 - 9 pertain to the general model (5.1).

Claim 5: Consider the model

$$\mu(x;\beta,\lambda,\rho) = [1 + \lambda x(\rho)\beta]^{1/\lambda},$$

where $x(\rho) \equiv (1, x_2, \ldots, x_J, x_{J+1}(\rho_{J+1}), \ldots, x_K(\rho_K))$. Suppose the scaled data are $c_0 y_t, c_{J+1} x_{t,J+1}, \ldots, c_K x_{tK}$ where $c_j > 0$. Then the relationships between the estimators using scaled data and unscaled data are

$$\lambda^+ = \hat{\lambda}; \quad \rho_j^+ = \hat{\rho}_j, \; j=J+1,\ldots,K;$$

$$\beta_1^+ = (c_0^{\hat{\lambda}} - 1)\hat{\lambda}^{-1} + c_0^{\hat{\lambda}}\{\hat{\beta}_1 + \hat{\beta}_{J+1}\hat{\rho}_{J+1}^{-1}(c_{J+1}^{\hat{\rho}_{J+1}} - 1) + \hat{\beta}_K \hat{\rho}_K^{-1}(c_K^{\hat{\rho}_K} - 1)\};$$

$$\beta_j^+ = c_0^{\hat{\lambda}}\hat{\beta}_j, \; j=2,\ldots,J; \quad \beta_j^+ = c_0^{\hat{\lambda}} c_j^{-\hat{\rho}_j}\hat{\beta}_j, \; j=J+1,\ldots,K.$$

Proof: This is similar to the proof of Claim 1. The first order condition for $\theta^+ \equiv (\beta^+, \lambda^+, \rho^+)$ is given by

$$\sum_{t=1}^{N} \nabla_\theta \mu(x_t^+;\theta^+)'(c_0 y_t - \mu(x_t^+;\theta^+)) \equiv 0; \qquad (a.13)$$

here, $x_t^+$ denotes the scaled regressors $x_t C$, where $C \equiv$ $\text{diag}(1,\ldots,1,c_{J+1},\ldots,c_K)$. Showing that the above choice of $\theta^+$ solves (a.13)

relies on the following relationships:

$$\nabla_{\beta}\mu(x_t^+;\theta^+) = c_0^{1-\hat{\lambda}}\nabla_{\beta}\mu(x_t;\hat{\theta}) \tag{a.14}$$

$$\nabla_{\lambda}\mu(x_t^+;\theta^+) = c_0\nabla_{\lambda}\mu(x_t;\hat{\theta}) + \tag{a.15}$$

$$\hat{\lambda}^{-2}c_0^{1-\hat{\lambda}}[1 + \hat{\lambda}x_t(\hat{\rho})\hat{\beta}]^{(1/\hat{\lambda})-1}[c_0^{\hat{\lambda}} + \hat{\lambda}\log(c_0)c_0^{\hat{\lambda}}(1 + \hat{\lambda}x_t(\hat{\rho})\hat{\beta}) - 1]$$

$$\nabla_{\rho_j}\mu(x_t^+,\theta^+) = c_0\nabla_{\rho_j}\mu(x_t,\hat{\theta}) + c_0\log(c_j)\hat{\beta}_j\nabla_{\beta_j}\mu(x_t,\hat{\theta}) \tag{a.16}$$

$$+ c_0[\{c_j^{-\hat{\rho}_j}/\hat{\rho}_j - 1/\hat{\rho}_j\}/\hat{\rho}_j - \log(c_j)/\hat{\rho}_j]\nabla_{\beta_1}\mu(x_t,\hat{\theta})$$

$$j=J+1,\ldots,K.$$

Equations (a.14), (a.15), and (a.16) show that $\nabla_{\theta}\mu_t(x_t^+,\theta^+)$ is a linear combination of $\nabla_{\theta}\mu(x_t,\hat{\theta})$. Because $\epsilon_t^+ = c_0\hat{\epsilon}_t$, it follows that

$$\sum_{t=1}^{N}\nabla_{\theta}\mu(x_t^+,\theta^+)'\epsilon_t^+ = 0,$$

which establishes (a.13). ■

Claim 6: The estimators $\hat{\beta}_2,\ldots,\hat{\beta}_J$ and the associated standard errors are invariant to scaling of only $x_{J+1},\ldots,x_K$.

Proof: In this case, $c_0 = 1$, so that $\beta_j^+ = \hat{\beta}_j$, $j=2,\ldots,J$ follows from Claim 5. Also, $\epsilon_t^+ = \hat{\epsilon}_t$, $\nabla_{\beta}\mu(x_t^+;\theta^+) = \mu(x_t;\hat{\theta})$, $\nabla_{\lambda}\mu(x_t^+;\theta^+) = \nabla_{\lambda}\mu(x_t;\hat{\theta})$, and

$$\nabla_{\rho_j}\mu(x_t^+,\theta^+) = \nabla_{\rho_j}\mu(x_t,\hat{\theta}) + \log(c_j)\hat{\beta}_j\nabla_{\beta_j}\mu(x_t,\hat{\theta})$$

$$+ [\{c_j^{-\hat{\rho}_j}/\hat{\rho}_j - 1/\hat{\rho}_j\}/\hat{\rho}_j - \log(c_j)/\hat{\rho}_j]\nabla_{\beta_1}\mu(x_t,\hat{\theta}),$$

$$j=J+1,\ldots,K.$$

As in the proof of Claim 2, the standard error of, say, $\beta_2^+ = \hat{\beta}_2$ depends only on $\epsilon_t^+$ and the residuals $r_{t2}^+$ from the regression

$$\nabla_{\beta_2}\mu_t^+ \text{ on } \nabla_{\beta_1}\mu_t^+, \nabla_{\beta_3}\mu_t^+, \ldots, \nabla_{\beta_K}\mu_t^+, \nabla_{\lambda}\mu_t^+, \nabla_{\rho}\mu_t^+, \quad t=1,\ldots,N.$$

Because of the above relationships, these residuals are independent of the

scale variables $c_{J+1},\ldots,c_K$. A similar argument in fact shows that the asymptotic variance associated of the subvector $(\hat{\beta}_2,\ldots,\hat{\beta}_J)$ is invariant with respect to the scale coefficients $c_{J+1},\ldots,c_K$. ■

Claim 7: The asymptotic variance of $(\hat{\lambda},\hat{\rho}_{J+1},\ldots,\hat{\rho}_K)$ is independent of $c_0,c_{J+1},\ldots,c_K$.

Proof: As in the proof of Claim 2, a mean value expansion shows that

$$\sqrt{N}[(\hat{\lambda},\hat{\rho}) - (\lambda,\rho)]' = \left(N^{-1}\sum_{t=1}^{N} \mathbf{r}_t^{o\prime}\mathbf{r}_t^{o}\right)^{-1} N^{-1/2}\sum_{t=1}^{N} \mathbf{r}_t^{o\prime}\epsilon_t^{o} + o_p(1),$$

where the 1x(1+K-J) vectors $\mathbf{r}_t^o$ are the residuals from the regression

$$\nabla_\lambda \mu_t^o,\ \nabla_\rho \mu_t^o \ \text{ on } \ \nabla_\beta \mu_t^o, \quad t=1,\ldots,N. \tag{a.17}$$

The gradients in (a.17) are evaluated at the scaled $\mathbf{x}_t$ and the scaled true coefficients $\theta^o$. Because $\epsilon_t^o = c_0\epsilon_t$, it suffices to show that $\mathbf{r}_t^o = c_0\mathbf{r}_t$, where $\mathbf{r}_t$ are the residuals from

$$\nabla_\lambda \mu_t,\ \nabla_\rho \mu_t \ \text{ on } \ \nabla_\beta \mu_t, \quad t=1,\ldots,N; \tag{a.18}$$

the gradients in (a.18) are evaluated at the unscaled quantities $\mathbf{x}_t$ and $\theta$. But the population analogs of (a.14)-(a.16) are

$$\nabla_\beta \mu_t^o = c_0^{1-\lambda}\nabla_\beta \mu_t$$

$$\nabla_\lambda \mu_t^o = c_0\nabla_\lambda \mu \mathbf{x}_t + \nabla_\beta \mu_t \mathbf{a}$$

$$\nabla_{\rho_j} \mu_t^o = c_0\nabla_{\rho_j} \mu_t + d_j\nabla_{\beta_j} \mu_t + f_j\nabla_{\beta_1} \mu_t, \quad j=J+1,\ldots,K,$$

for constants $d_j$ and $f_j$; it immediately follows that $\mathbf{r}_t^o = c_0\mathbf{r}_t$, and this shows that the asymptotic variance of $(\hat{\lambda},\hat{\rho})$ is scale invariant. ■

Claim 8: LM tests for exclusion restrictions are independent of $c_0,c_{J+1},\ldots,c_K$.

Proof: The unrestricted model is

$$\mu(\mathbf{x},\mathbf{z};\beta,\lambda,\rho,\delta) = [1 + \lambda\mathbf{x}(\rho)\beta + \lambda\mathbf{z}\delta]^{1/\lambda},$$

and the null hypothesis is $H_0$: $\delta = 0$. The relationships between the restricted gradients for the scaled and unscaled data are given by (a.14)-(a.16) for the parameters $\beta$, $\lambda$, and $\rho$. The gradient for $\delta$ evaluated at $\delta = 0$ is easily seen to satisfy

$$\nabla_\delta \mu(\mathbf{x}_t^+;\theta^+,0) = c_0^{1-\hat{\lambda}} \nabla_\delta \mu(\mathbf{x}_t;\hat{\theta},0) = c_0^{1-\hat{\lambda}} [1 + \hat{\lambda}\mathbf{x}_t(\hat{\rho})\hat{\beta}]^{\{(1/\hat{\lambda}) - 1\}} \mathbf{z}_t$$

Thus, the r-squared from the regression

$$\epsilon_t^+ \text{ on } \nabla_\beta \mu(\mathbf{x}_t^+;\theta^+,0),\ \nabla_\lambda \mu(\mathbf{x}_t^+;\theta^+,0),\ \nabla_\rho \mu(\mathbf{x}_t^+;\theta^+,0),\ \nabla_\delta \mu(\mathbf{x}_t^+;\theta^+,0)$$

is independent of $c_0, c_{J+1}, \ldots, c_K$. A similar argument establishes that the robust LM test is also invariant. ■

Claim 9: In the model (6.10) with intial scale estimates $\hat{\nu}$, $\hat{\eta}$, (6.16) is a consistent estimator of the asymptotic variance of $\hat{\theta}$. Moreover, the submatrix corresponding to $(\hat{\lambda},\hat{\rho})$ is unaffected by the asymptotic variance of $(\hat{\nu},\hat{\eta})$.

Proof: As in the proof of Claim 4, a standard mean value expansion shows that

$$\sqrt{N}(\hat{\theta} - \theta) = \left(N^{-1}\sum_{t=1}^{N} \nabla_\theta \mu_t' \nabla_\theta \mu_t\right)^{-1} N^{-1/2} \sum_{t=1}^{N} \nabla_\theta \mu_t' \epsilon_t \qquad (a.19)$$

$$- \left(N^{-1}\sum_{t=1}^{N} \nabla_\theta \mu_t' \nabla_\theta \mu_t\right)^{-1} \left(N^{-1}\sum_{t=1}^{N} \nabla_\theta \mu_t' \nabla_\pi \mu_t\right) N^{-1/2} \sum_{t=1}^{N} g_t' + o_p(1),$$

where $\theta \equiv (\beta',\lambda,\rho')'$, $\pi \equiv (\nu,\eta')'$, and $g_t$ is the 1x(1+K-J) vector

$$g_t \equiv [\nu\log(y_t/\nu), \eta_{J+1}\log(x_{t,J+1}/\eta_{J+1}), \ldots, \eta_K \log(x_{tK}/\eta_K)].$$

(Note that $E(g_t) = 0$.) Thus,

$$\sqrt{N}(\hat{\theta} - \theta) = A_N^{-1}[I_P|-C_N]\left[N^{-1/2}\sum_{t=1}^{N}\nabla_\theta\mu_t\epsilon_t,\ -N^{-1/2}\sum_{t=1}^{N}g_t\right]' + o_p(1)$$

where

$$A_N \equiv N^{-1}\sum_{t=1}^{N}E[\nabla_\theta\mu_t'\nabla_\theta\mu_t]$$

$$C_N \equiv N^{-1}\sum_{t=1}^{N}E[\nabla_\theta\mu_t'\nabla_\pi\mu_t].$$

This is written concisely as

$$\sqrt{N}(\hat{\theta} - \theta) = A_N^{-1}[I_P|-C_N]N^{-1/2}\sum_{t=1}^{N}s_t' + o_p(1)$$

where $s_t$ is the 1x(P+1+K-J) vector $s_t \equiv (\nabla_\theta\mu_t\epsilon_t, g_t)$. Thus, the asymptotic variance of $\sqrt{N}(\hat{\theta} - \theta)$ is

$$A_N^{-1}[I_P|-C_N]\left[N^{-1}\sum_{t=1}^{N}E(s_t's_t)\right][I_P|-C_N]'A_N^{-1}.$$

Replacing unknown expecations by their sample counterparts, and the unknown parameters by the estimates $(\hat{\theta},\hat{\pi})$, yields (6.16) multiplied by N; the asymptotic variance of $\hat{\theta}$ is obtained by dividing $AV[\sqrt{N}(\hat{\theta} - \theta)]$ by N. This completes the first part of the assertion.

To establish the second part, we show that the elements in the last 1+(K-J) rows of

$$\left[\sum_{t=1}^{N}\nabla_\theta\mu_t'\nabla_\theta\mu_t\right]^{-1}\left[\sum_{t=1}^{N}\nabla_\theta\mu_t'\nabla_\pi\mu_t\right] \qquad (a.20)$$

(those corresponding to $(\lambda,\rho)$) are identically zero. Let $R_t$ be the (1+K-J)xK matrix of residuals from the matrix regression

$$\nabla_\lambda\mu_t,\ \nabla_\rho\mu_t \quad \text{on} \quad \nabla_\beta\mu_t;$$

then it suffices to show that

$$\sum_{t=1}^{N} R_t' \nabla_\pi \mu_t = 0. \qquad \text{(a.21)}$$

But (a.21) holds if $\nabla_\pi \mu_t$ is a linear combination of $\nabla_\beta \mu_t$, and this is seen to be the case from (6.14) provided that $x_t$ contains a constant. The sample counterpart of this argument shows that (6.16) is numerically identical to the robust variance estimate for the subvector $(\hat{\lambda}, \hat{\rho})$. ■