Non-nested Hypothesis Testing: An Overview

M. Hashem Pesaran and Melvyn Weeks
Faculty of Economics and Politics
University of Cambridge

September 1999

1. Introduction

This chapter focuses on the hypotheses testing problem when the hypotheses or models under consideration are "non-nested" or belong to "separate" families of distributions, in the sense that none of the individual models may be obtained from the remaining either by imposition of parameter restrictions or through a limiting process[1]. In econometric analysis non-nested models arise naturally when rival economic theories are used to explain the same phenomenon such as unemployment, inflation or output growth. Typical examples from the economics literature are Keynesian and new classical explanations of unemployment, structural and monetary theories of inflation, alternative theories of investment, and endogenous and exogenous theories of growth.[2] Non-nested models could also arise when alternative functional specifications are considered such as multinomial probit and logit distribution functions used in the qualitative choice literature, exponential and power utility functions used in the asset pricing models, and a variety of non-nested specifications considered in the empirical analysis of income and wealth distributions. Finally, even starting from the same theoretical paradigm, it is possible for different investigators to arrive at different models if they adopt different conditioning or follow different paths to a more parsimonious model using the general-to-specific specification search methodology, advocated, for example by Hendry (1993).

The concept of an econometric model is discussed in Section 2, where a distinction is made between conditional and unconditional models. This is an important distinction since most applied work in econometrics takes place within a modelling framework where the behaviour of one or more "endogenous" variables is often explained *conditional* on a set of "exogenous" variables. This discussion also highlights the importance of conditioning in the process of model evaluation. Examples of non-nested models are given in Section 3. Section 4 discusses the differences that lie

[1] Therefore our focus is distinct from Chow (1981) who, in examining a similar problem, assumes that the set of models under consideration contains a general model from which all other competing models may be obtained by the imposition of suitable parameter restrictions.

[2] See, for example, Friedman and Meiselman (1963) on alternative consumption models, Barro (1977), Pesaran (1982b) and McAleer, Pesaran, and Bera (1990) on alternative explanations of the unemployment rate, Jorgenson and Siebert (1968), Dixit and Pindyck (1994) and Bernanke, Bohn, and Reiss (1988) on alternative models of investment behaviour, and McAleer, Fisher, and Volker (1982) and Smith and Smyth (1991) on non-nested money demand functions.

behind model selection and hypotheses testing. Although this chapter is primarily concerned with hypotheses testing involving non-nested models, a discussion of the differences and similarities of the two approaches to model evaluation can serve an important pedagogic purpose in clarifying the conditions under which one approach rather than the other could be appropriate.

The literature on non-nested hypothesis testing in statistics was pioneered by the seminal contributions of Cox (1961), Cox (1962) and Atkinson (1970), and was subsequently applied to econometric models by Pesaran (1974) and Pesaran and Deaton (1978). The analysis of non-nested regression models was further considered by Davidson and MacKinnon (1981), Fisher and McAleer (1981), Dastoor (1983), Deaton (1982), Sawyer (1983), Gourieroux, Monfort, and Trognon (1983), and Godfrey and Pesaran (1983)[3]. .This literature is reviewed in Section 5 where we examine a number of alternative approaches to testing non-nested hypotheses, including the encompassing approach advanced by Mizon and Richard (1986), Gourieroux and Monfort (1995) and Smith (1993).

Generally speaking, two models, say H_f and H_g, are said to be non-nested if it is not possible to derive H_f (or H_g) from the other model either by means of an exact set of parametric restrictions or as a result of a limiting process. But for many purposes a more rigorous definition is needed. Section 6 examines this issue and focuses on the Kullback-Leibler divergence measure which has played a pivotal role in the development of a number of non-nested test statistics. The Vuong approach to model selection, viewed as a hypothesis testing problem is also discussed in this section. (see Vuong (1989)). Section 7 deals with the practical problems involved in the implementation of the Cox procedure. Apart from a few exceptions, the centring of the log-likelihood ratio statistic required to construct the Cox statistic, will involve finding an estimate of the Kullback-Leibler measure of closeness of the alternative to the null hypothesis, which in most cases is not easy to compute using analytical techniques. Subsequently we explore two methods which circumvent the problem. First, following work by Pesaran and Pesaran (1993), we examine the simulation approach which provides a consistent estimator of the KLIC measure. However, since this approach is predicated upon the adherence to a classical testing framework, we also examine the use of a parametric bootstrap approach. Whereas the use of simulation facilitates the construction of a pivotal test statistic with an asymptotically well defined limiting distribution, the bootstrap procedure effectively replaces the theoretical distribution with the empirical distribution function. We also discuss the use of pivotal bootstrap statistics for testing non-nested models.

[3]An excellent survey article on non-nested hypothesis testing can be found in Gourieroux and Monfort (1994).

2. Models and Their Specification

Suppose the focus of the analysis is to consider the behaviour of the $n \times 1$ vector of random variables $\mathbf{w}_t = (\mathbf{w}_{1t}, \mathbf{w}_{2t}, ..., \mathbf{w}_{nt})'$ observed over the period $t = 1, 2, ..., T$. A model of $\mathbf{w}_t$, indexed by $\mathfrak{M}_i$, is defined by the joint probability distribution function (p.d.f.) of the observations

$$\begin{aligned} \mathbf{W} &= (\mathbf{w}_1', \mathbf{w}_2', ..., \mathbf{w}_T')' \\ \mathfrak{M}_i &: f_i(\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_T | \mathbf{w}_0, \boldsymbol{\varphi}_i) = f_i(\mathbf{W}|\mathbf{w}_0, \boldsymbol{\varphi}_i), \qquad i = 1, 2, ..., m, \end{aligned} \tag{2.1}$$

where $f_i(.)$ is the probability density function of the model (hypothesis) $\mathfrak{M}_i$, and $\boldsymbol{\varphi}_i$ is a $p_i \times 1$ vector of unknown parameters associated with model $\mathfrak{M}_i$.[4]

The models characterised by $f_i(\mathbf{W}|\mathbf{w}_0, \boldsymbol{\varphi}_i)$ are *unconditional* in the sense that probability distribution of $\mathbf{w}_t$ is fully specified in terms of some initial values, $\mathbf{w}_0$, and for a given value of $\boldsymbol{\varphi}_i$. In econometrics the interest often centres on conditional models, where a vector of "endogenous" variables, $\mathbf{y}_t$, is explained (or modelled) *conditional* on a set of "exogenous", variables, $\mathbf{x}_t$. Such conditional models can be derived from (2.1) by noting that

$$\begin{aligned} & f_i(\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_T | \mathbf{w}_0, \boldsymbol{\varphi}_i) \\ = \; & f_i(\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_T | \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T, \boldsymbol{\psi}(\boldsymbol{\varphi}_i)) \times f_i(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T | \mathbf{w}_0, \boldsymbol{\kappa}(\boldsymbol{\varphi}_i)), \end{aligned} \tag{2.2}$$

where $\mathbf{w}_t = (\mathbf{y}_t', \mathbf{x}_t')$. The unconditional model $\mathfrak{M}_i$ is decomposed into a conditional model of $\mathbf{y}_t$ given $\mathbf{x}_t$ and a marginal model of $\mathbf{x}_t$. Denoting the former by $\mathfrak{M}_{i,y|x}$ we have

$$\mathfrak{M}_{i,y|x} : f_i(\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_T | \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T, \mathbf{w}_0, \boldsymbol{\psi}(\boldsymbol{\varphi}_i)) = f_i(\mathbf{Y}|\mathbf{X}, \mathbf{w}_0, \boldsymbol{\psi}(\boldsymbol{\varphi}_i)), \tag{2.3}$$

where $\mathbf{Y} = (\mathbf{y}_1', \mathbf{y}_2', ..., \mathbf{y}_T')'$ and $\mathbf{X} = (\mathbf{x}_1', \mathbf{x}_2', ..., \mathbf{x}_T')'$.

Confining attention to the analysis and comparison of conditional models is valid only if the variations in the parameters of the marginal model, $\boldsymbol{\kappa}(\boldsymbol{\varphi}_i)$, does not induce changes in the parameters of the conditional model, $\boldsymbol{\psi}(\boldsymbol{\varphi}_i)$. Namely $\partial \boldsymbol{\psi}(\boldsymbol{\varphi}_i)/\partial' \boldsymbol{\kappa}(\boldsymbol{\varphi}_i) = 0$. When this condition holds it is said that $\mathbf{x}_t$ is *weakly exogenous* for $\boldsymbol{\psi}_i$. The parameters of the conditional model, $\boldsymbol{\psi}_i = \boldsymbol{\psi}(\boldsymbol{\varphi}_i)$, are often referred as the *parameters of interest*.[5]

The conditional models $\mathfrak{M}_{i,y|\mathbf{x}}$ $i = 1, 2, ..., m$ all are based on the same conditioning variables, $\mathbf{x}_t$, and differ only insofar as they are based upon different p.d.fs. We may introduce an alternative set of models which share the same p.d.fs but differ with respect to the inclusion of exogenous variables. For any model, $\mathfrak{M}_i$ we may partition the set of exogenous variables $\mathbf{x}_t$ according to a simple included/excluded

[4] In cases where one or more elements of $\mathbf{z}_t$ are discrete, as in probit or Tobit specifications cumulative probabality distribution functions can be used instead of probability density functions.

[5] See Engle, Hendry and Richard (1983).

dichotomy. Therefore $\mathbf{x}_t = (\mathbf{x}_{it}', \mathbf{x}_{it}^{*\prime})'$ writes the set of exogenous variables according to a subset $\mathbf{x}_{it}$ which are included in model $\mathfrak{M}_i$, and a subset $\mathbf{x}_{it}^*$ which are excluded. We may then write

$$\begin{aligned} & f_i(\mathbf{Y}|\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_T, \mathbf{w}_0, \boldsymbol{\varphi}_i) \\ = \; & f_i(\mathbf{Y}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, ...\mathbf{x}_{iT}, \mathbf{x}_{i1}^*, \mathbf{x}_{i2}^*, ..., \mathbf{x}_{iT}^*, \mathbf{w}_0, \boldsymbol{\varphi}_i) \\ = \; & f_i(\mathbf{Y}|\mathbf{X}_i, \mathbf{w}_0, \boldsymbol{\psi}_i(\boldsymbol{\varphi}_i)) \times f_i(\mathbf{X}_i^*|\mathbf{X}_i, \mathbf{w}_0, \mathbf{c}_i(\boldsymbol{\varphi}_i)) \end{aligned}$$

where $\mathbf{X}_i = (\mathbf{x}_{i1}', \mathbf{x}_{i2}', ..., \mathbf{x}_{iT}')'$ and $\mathbf{X}_i^* = (\mathbf{x}_{i1}^{*\prime}, \mathbf{x}_{i2}^{*\prime}, ..., \mathbf{x}_{iT}^{*\prime})'$. As noted above in the case of models differentiated solely by different p.d.fs, a comparison of models based upon the partition of $\mathbf{x}_t$ into $\mathbf{x}_{it}$ and $\mathbf{x}_{it}^*$ should be preceded by determining whether $\partial\boldsymbol{\psi}_i(\boldsymbol{\varphi}_i)/\partial\mathbf{c}_i'(\boldsymbol{\varphi}_i) = 0$.

The above set up allows consideration of rival models that could differ in the conditioning set of variables, $\{\mathbf{x}_{it},\ i = 1, 2, ..., m\}$ and/or the functional form of their underlying probability distribution functions, $\{f_i(\cdot),\ i = 1, 2, ..., m\}$. In much of this chapter we will be concerned with two rival (conditional) models and for notational convenience we denote them by

$$H_f \;\; : \;\; \mathcal{F}_\theta = \{f(\mathbf{y}_t|\mathbf{x}_t, \Omega_{t-1}; \boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta\}, \tag{2.4}$$

$$H_g \;\; : \;\; \mathcal{F}_\gamma = \{g(\mathbf{y}_t|\mathbf{z}_t, \Omega_{t-1}; \boldsymbol{\gamma}),\ \boldsymbol{\gamma} \in \Gamma\}, \tag{2.5}$$

where Ω_{t-1} denotes the set of all past observations on $\mathbf{y}, \mathbf{x}$ and $\mathbf{z}$, $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ are respectively k_f and k_g vectors of unknown parameters belonging to the non-empty compact sets Θ and Γ, and where $\mathbf{x}$ and $\mathbf{z}$ represent the conditioning variables. For the sake of notational simplicity we shall also often use $f_t(\boldsymbol{\theta})$ and $g_t(\boldsymbol{\gamma})$ in place of $f(\mathbf{y}_t|\mathbf{x}_t, \Omega_{t-1}; \boldsymbol{\theta})$ and $g(\mathbf{y}_t|\mathbf{z}_t, \Omega_{t-1}; \boldsymbol{\gamma})$, respectively.

Now given the observations $(\mathbf{y}_t, \mathbf{x}_t, \mathbf{z}_t,\ t = 1, 2, ..., T)$ and conditional on the initial values $\mathbf{w}_0$, the maximum likelihood (ML) estimators of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ are given by

$$\widehat{\boldsymbol{\theta}}_T = \underset{\boldsymbol{\theta}\in\Theta}{Arg\max} L_f(\boldsymbol{\theta}), \quad \hat{\boldsymbol{\gamma}}_T = \underset{\boldsymbol{\gamma}\in\Gamma}{Arg\max} L_g(\boldsymbol{\gamma}), \tag{2.6}$$

where the respective log-likelihood functions are given by:

$$L_f(\boldsymbol{\theta}) = \sum_{t=1}^{T} \ln f_t(\boldsymbol{\theta}), \quad L_g(\boldsymbol{\gamma}) = \sum_{t=1}^{T} \ln g_t(\boldsymbol{\gamma}). \tag{2.7}$$

Throughout we shall assume that the conditional densities $f_t(\boldsymbol{\theta})$ and $g_t(\boldsymbol{\gamma})$ satisfy the usual requilarity conditions as set out, for example, in White (1982) and Smith (1993), needed to ensure that $\hat{\boldsymbol{\theta}}_T$ and $\hat{\boldsymbol{\gamma}}_T$ have asymptotically normal limiting distributions under the Data Generating Process (DGP). We allow the DGP to differ from H_f and H_g, and denote it by H_h; thus admitting the possibility that both H_f and H_g could be misspecified and that both are likely to be rejected in practice.

In this setting $\hat{\boldsymbol{\theta}}_T$ and $\hat{\boldsymbol{\gamma}}_T$ are referred to as quasi-ML estimators and their probability limits under H_h, which we denote by $\boldsymbol{\theta}_{h*}$ and $\boldsymbol{\gamma}_{h*}$ respectively, are known as (asymptotic) pseudo-true values. These pseudo-true values are defined by

$$\boldsymbol{\theta}_{h*} = \underset{\boldsymbol{\theta}\in\Theta}{Arg\max}\ E_h\{T^{-1}L_f(\boldsymbol{\theta})\},\ \ \boldsymbol{\gamma}_{h*} = \underset{\boldsymbol{\gamma}\in\Gamma}{Arg\max}\ E_h\{T^{-1}L_g(\boldsymbol{\gamma})\}, \tag{2.8}$$

where $E_h(\cdot)$ denotes expectations are taken under H_h. In the case where $\mathbf{w}_t$ follows a strictly stationary process, (2.8) simplifies to

$$\boldsymbol{\theta}_{h*} = \underset{\boldsymbol{\theta}\in\Theta}{Arg\max}\ E_h\{\ln f_t(\boldsymbol{\theta})\},\quad \boldsymbol{\gamma}_{h*} = \underset{\boldsymbol{\gamma}\in\Gamma}{Arg\max}\ E_h\{\ln g_t(\boldsymbol{\gamma})\}. \tag{2.9}$$

To ensure global identifiability of the pseudo-true values, it will be assumed that $\boldsymbol{\theta}_{f*}$ and $\boldsymbol{\gamma}_{f*}$ provide *unique* maxima of $E_h\{T^{-1}L_f(\boldsymbol{\theta})\}$ and $E_h\{T^{-1}L_g(\boldsymbol{\gamma})\}$, respectively. Clearly, under H_f, namely assuming H_f is the DGP, we have $\boldsymbol{\theta}_{f*}$= $\boldsymbol{\theta}_0$, and $\boldsymbol{\gamma}_{f*} = \boldsymbol{\gamma}_*(\boldsymbol{\theta}_0)$ where $\boldsymbol{\theta}_0$ is the "true" value of $\boldsymbol{\theta}$ under H_f. Similarly, under H_g we have $\boldsymbol{\gamma}_{g*}$= $\boldsymbol{\gamma}_0$, and $\boldsymbol{\theta}_{g*}$= $\boldsymbol{\theta}_*(\boldsymbol{\gamma}_0)$ with $\boldsymbol{\gamma}_0$ denoting the "true" value of $\boldsymbol{\gamma}$ under H_g. The functions $\boldsymbol{\gamma}_*(\boldsymbol{\theta}_0)$,and $\boldsymbol{\theta}_*(\boldsymbol{\gamma}_0)$ that relate the parameters of the two models under consideration are called the *binding* functions. These functions do not involve the true model, H_h, and only depend on the models H_f and H_g that are under consideration. As we shall see later a formal definition of encompassing is given in terms of the pseudo true values, $\boldsymbol{\theta}_{h*}$ and $\boldsymbol{\gamma}_{h*}$, and the binding functions $\boldsymbol{\gamma}_*(\boldsymbol{\theta}_0)$, and $\boldsymbol{\theta}_*(\boldsymbol{\gamma}_0)$.

Before proceeding further it would be instructive to consider some examples of non-nested models from the literature.

3. Examples of Non-Nested Models

We start with examples of unconditional non-nested models. One such example, originally discussed by Cox (1961) is that of testing a log-normal versus an exponential distribution.

$$\begin{aligned} H_f\ &:\ f(y_t|\boldsymbol{\theta}) = f_t(\boldsymbol{\theta}) = y_t^{-1}(2\pi\theta_2)^{-1/2}\exp\left\{-\frac{(\ln y_t-\theta_1)^2}{2\theta_2}\right\},\ \infty>\theta_2>0,\ y_t>0 \\ H_g\ &:\ g(y_t|\boldsymbol{\gamma}) = g_t(\boldsymbol{\gamma}) = \boldsymbol{\gamma}^{-1}\exp(-y_t/\boldsymbol{\gamma}),\ \ \boldsymbol{\gamma}>0,\ y_t>0. \end{aligned}$$

These hypotheses (models) are globally non-nested, in the sense that neither can be obtained from the other either by means of suitable parametric restrictions or by a limiting process.[6] Under H_f the pseudo-true value of $\boldsymbol{\gamma}$, denoted by $\boldsymbol{\gamma}_{f*}$ is obtained by solving the following maximization problem

$$\boldsymbol{\gamma}_{f*} = \underset{\boldsymbol{\gamma}>0}{Arg\max}\ E_f\{\ln g_t(\boldsymbol{\gamma})\}.$$

[6] A formalization of the concept of globally non-nested models can be found in Pesaran (1987). Also see Section 6.

But[7]

$$E_f\{\ln g_t(\boldsymbol{\gamma})\} = -\ln \boldsymbol{\gamma} - E_f(y_t)/\boldsymbol{\gamma} = -\ln \boldsymbol{\gamma} - \exp(\theta_1 + 0.5\theta_2)/\boldsymbol{\gamma},$$

which yields

$$\boldsymbol{\gamma}_{f*} = \boldsymbol{\gamma}_*(\boldsymbol{\theta}_0) = \exp(\theta_{10} + 0.5\theta_{20}).$$

Similarly, under H_g we have[8]

$$\theta_1^*(\lambda_0) = \ln \boldsymbol{\gamma}_0 - 0.5772, \qquad \theta_2^*(\boldsymbol{\gamma}_0) = 1.6449.$$

Other examples of non-nested unconditional models include log-normal versus Weibull, Pereira (1984) and log-normal versus gamma distribution, Pesaran (1987).

The most prominent example of conditional non-nested models is linear normal regression models with "rival" sets of conditioning variables. As an example consider the following regression models:

$$H_f \quad : \quad y_t = \boldsymbol{\alpha}'\mathbf{x}_t + u_{tf}, \quad u_{tf} \backsim N(0, \sigma^2), \quad \infty > \sigma^2 > 0, \tag{3.1}$$

$$H_g \quad : \quad y_t = \boldsymbol{\beta}'\mathbf{z}_t + u_{tg}, \quad u_{tg} \backsim N(0, \omega^2), \quad \infty > \omega^2 > 0. \tag{3.2}$$

The conditional probability density associated with these regression models are given by

$$H_f : \ f(y_t|\mathbf{x}_t; \boldsymbol{\theta}) = (2\pi\sigma^2)^{-1/2} \exp\left\{\frac{-1}{2\sigma^2}(y_t - \boldsymbol{\alpha}'\mathbf{x}_t)^2\right\}, \tag{3.3}$$

$$H_g : \ g(y_t|\mathbf{z}_t; \boldsymbol{\theta}) = (2\pi\omega^2)^{-1/2} \exp\left\{\frac{-1}{2\omega^2}(y_t - \boldsymbol{\beta}'\mathbf{z}_t)^2\right\}, \tag{3.4}$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \sigma^2)'$, and $\boldsymbol{\gamma} = (\boldsymbol{\beta}', \omega^2)'$. These regression models are non-nested if it is not possible to write $\mathbf{x}_t$ as an exact linear function of $\mathbf{z}_t$ and *vice versa*, or more formally if $\mathbf{x}_t \nsubseteq \mathbf{z}_t$ *and* $\mathbf{z}_t \nsubseteq \mathbf{x}_t$. Model H_f is said to be nested in H_g if $\mathbf{x}_t \subset \mathbf{z}_t$ *and* $\mathbf{z}_t \nsubseteq \mathbf{x}_t$. The two models are observationally equivalent if $\mathbf{x}_t \subset \mathbf{z}_t$ *and* $\mathbf{z}_t \subset \mathbf{x}_t$. Suppose now that neither of these regression models is true and the DGP is given by

$$H_h : \ y_t = \boldsymbol{\delta}'\mathbf{w}_t + u_{th}, \quad u_{th} \backsim N(0, \upsilon^2), \quad \infty > \upsilon^2 > 0. \tag{3.5}$$

It is then easily seen that conditional on $\{\mathbf{x}_t, \mathbf{z}_t, \mathbf{w}_t, \ t = 1, 2, ..., T\}$

$$E_h\left\{T^{-1}L_f(\boldsymbol{\theta})\right\} = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{\upsilon^2}{2\sigma^2} - \frac{\boldsymbol{\delta}'\hat{\Sigma}_{ww}\boldsymbol{\delta} - 2\boldsymbol{\delta}'\hat{\Sigma}_{wx}\alpha + \boldsymbol{\alpha}'\hat{\Sigma}_{xx}\alpha}{2\sigma^2},$$

[7] Note that under H_f, $E(y_t) = E\{\exp(\ln y_t)\} = \exp(\theta_1 + 0.5\theta_2)$.

[8] See, Pesaran (1984, p. 249-50).

where

$$\hat{\Sigma}_{ww} = T^{-1}\sum_{t=1}^{T}\mathbf{w}_t\mathbf{w}_t',\ \hat{\Sigma}_{xx} = T^{-1}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t',\ \hat{\Sigma}_{wx} = T^{-1}\sum_{t=1}^{T}\mathbf{w}_t\mathbf{x}_t'.$$

Maximizing $E_h\{T^{-1}L_f(\boldsymbol{\theta})\}$ with respect to $\boldsymbol{\theta}$ now yields the conditional pseudo-true values:

$$\boldsymbol{\theta}_{h*} = \begin{pmatrix} \alpha_{h*} \\ \sigma^2_{h*} \end{pmatrix} = \begin{pmatrix} \hat{\Sigma}_{xx}^{-1}\hat{\Sigma}_{xw}\boldsymbol{\delta} \\ \upsilon^2 + \boldsymbol{\delta}'(\hat{\Sigma}_{ww} - \hat{\Sigma}_{wx}\hat{\Sigma}_{xx}^{-1}\hat{\Sigma}_{xw})\boldsymbol{\delta} \end{pmatrix}. \tag{3.6}$$

Similarly,

$$\boldsymbol{\gamma}_{h*} = \begin{pmatrix} \beta_{h*} \\ \omega^2_{h*} \end{pmatrix} = \begin{pmatrix} \hat{\Sigma}_{zz}^{-1}\hat{\Sigma}_{zw}\boldsymbol{\delta} \\ \upsilon^2 + \boldsymbol{\delta}'(\hat{\Sigma}_{ww} - \hat{\Sigma}_{wz}\hat{\Sigma}_{zz}^{-1}\hat{\Sigma}_{zw})\boldsymbol{\delta} \end{pmatrix}, \tag{3.7}$$

where

$$\hat{\Sigma}_{zz} = T^{-1}\sum_{t=1}^{T}\mathbf{z}_t\mathbf{z}_t',\ \hat{\Sigma}_{wz} = T^{-1}\sum_{t=1}^{T}\mathbf{w}_t\mathbf{z}_t'.$$

When the regressors are stationary, the unconditional counterparts of the above pseudo-true values can be obtained by replacing $\hat{\Sigma}_{ww}$, $\hat{\Sigma}_{xx}$, $\hat{\Sigma}_{wx}$ etc. by their population values, namely $\Sigma_{ww} = E(\mathbf{w}_t\mathbf{w}_t')$, $\Sigma_{xx} = E(\mathbf{x}_t\mathbf{x}_t')$, $\Sigma_{wx} = E(\mathbf{w}_t\mathbf{x}_t')$ etc.

Other examples of non-nested regression models include models with endogenous regressors estimated by instrumental variables (see, for example, Ericsson (1983) and Godfrey (1983)), non-nested non-linear regression models and regression models where the left hand side variables of the rival regressions are known transformations of a dependent variable of interest. One important instance of this last example is the problem of testing linear versus log-linear regression models and *vice versa.*[9] More generally we may have

$$H_f:\ f(y_t) = \boldsymbol{\alpha}'\mathbf{x}_t + u_{tf}, \quad u_{tf} \backsim N(0,\sigma^2),\ \ \infty > \sigma^2 > 0,$$

$$H_g:\ g(y_t) = \boldsymbol{\beta}'\mathbf{z}_t + u_{tg}, \quad u_{tg} \backsim N(0,\omega^2),\ \ \infty > \omega^2 > 0,$$

where $f(y_t)$ and $g(y_t)$ are known one-to-one functions of y_t. Within this more general regression framework testing a linear versus a log-linear model is characterized by $f(y_t) = y_t$ and $g(y_t) = \ln(y_t)$; a ratio model versus a log-linear model by $f(y_t) = y_t/q_t$ and $g(y_t) = \ln(y_t)$, where q_t is an observed regressor, and a ratio versus a linear model

[9] There is substantial literature on non-nested tests of linear versus log-linear regression models. Earlier studies include Aneuryn-Evans and Deaton (1980), Godfrey and Wickens (1981) and Davidson and MacKinnon (1985). In a more recent study Pesaran and Pesaran (1995) have examined the properties of a simulation-based variant of the Cox test.

by $f(y_t) = y_t/q_t$ and $g(y_t) = y_t$. For example, in analysis of aggregate consumption a choice needs to be made between a linear and a log-linear specification of the aggregate consumption on the one hand, and between a log-linear and a saving rate formulation on the other hand. The testing problem is further complicated here due to the linear transformations of the dependent variable, and additional restrictions are required if the existence of pseudo-true values in the case of these models are to be ensured. For example, suitable truncation restrictions need to be imposed on the errors of the linear model when it is tested against a log-linear alternative.

Other examples where specification of an appropriate error structure plays an important role in empirical analysis include discrete choice and duration models used in microeconometric research. Although the analyst may utilise both prior knowledge and theory to select an appropriate set of regressors, there is generally little guidance in terms of the most appropriate probability distribution. Non-nested hypothesis testing is particularly relevant to microeconometric research where the same set of regressors are often used to explain individual decisions but based on different functional distributions, such as multinomial probit and logit specifications in the analysis of discrete choice, exponential and Weibull distributions in the analysis of duration data. In the simple case of a probit (H_f) versus a logit model (H_g) we have

$$H_f \quad : \quad \Pr(y_t = 1) = \Phi(\boldsymbol{\theta}'\mathbf{x}_t) = \int_{-\infty}^{\boldsymbol{\theta}'\mathbf{x}_t} \frac{1}{\sqrt{2\pi}} \exp\left\{-\tfrac{1}{2}v^2\right\} dv \tag{3.8}$$

$$H_g \quad : \quad \Pr(y_t = 1) = \Lambda(\boldsymbol{\gamma}'\mathbf{z}_t) = \frac{e^{\boldsymbol{\gamma}'\mathbf{z}_t}}{1 + e^{\boldsymbol{\gamma}'\mathbf{z}_t}} \tag{3.9}$$

where y_t, $t = 1, 2, ..., T$ are independently distributed binary random variables taking the value of 1 or 0. In practice the two sets of regressors $\mathbf{x}_t$ used in the probit and logit specifications are likely to be identical, and it is only the form of the distribution functions that separate the two models. Other functional forms can also be entertained. Suppose, for example, that the true DGP for this simple discrete choice problem is given by the probability distribution function $H(\boldsymbol{\delta}'\mathbf{x}_t)$, then pseudo-true values for $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ can be obtained as functions of $\boldsymbol{\delta}$, but only in an implicit form. We first note that the log-likelihood function under H_f, for example, is given by

$$L_f(\boldsymbol{\theta}) = \sum_{t=1}^{T} y_t \log\left[\Phi(\boldsymbol{\theta}'\mathbf{x}_t)\right] + \sum_{t=1}^{T} (1 - y_t) \log\left[1 - \Phi(\boldsymbol{\theta}'\mathbf{x}_t)\right],$$

and hence under the assumed DGP we have

$$E_h\left\{T^{-1}L_f(\boldsymbol{\theta})\right\} = T^{-1}\sum_{t=1}^{T} H(\boldsymbol{\delta}'\mathbf{x}_t) \log\left[\Phi(\boldsymbol{\theta}'\mathbf{x}_t)\right] + T^{-1}\sum_{t=1}^{T}\left[1 - H(\boldsymbol{\delta}'\mathbf{x}_t)\right] \log\left[1 - \Phi(\boldsymbol{\theta}'\mathbf{x}_t)\right].$$

Therefore, the pseudo-true value of $\boldsymbol{\theta}$, namely $\boldsymbol{\theta}_*(\boldsymbol{\delta})$ or simply $\boldsymbol{\theta}_*$, satisfies the following equation

$$T^{-1}\sum_{t=1}^{T}\mathbf{x}_t\phi(\boldsymbol{\theta}'_*\mathbf{x}_t)\left\{\frac{H(\boldsymbol{\delta}'\mathbf{x}_t)}{\Phi(\boldsymbol{\theta}'_*\mathbf{x}_t)}-\frac{1-H(\boldsymbol{\delta}'\mathbf{x}_t)}{1-\Phi(\boldsymbol{\theta}'_*\mathbf{x}_t)}\right\}=0,$$

where $\phi(\boldsymbol{\theta}'_*\mathbf{x}_t)=(2\pi)^{-1/2}\exp\left[\frac{-1}{2}(\boldsymbol{\theta}'_*\mathbf{x}_t)^2\right]$. Using results in Amemiya (1985), pp. 271-2 it is easily established that the solution of $\boldsymbol{\theta}_*$ in terms of $\boldsymbol{\delta}$ is in fact unique, and $\boldsymbol{\theta}_*=\delta$ if and only if $\Phi(\cdot)=H(\cdot)$. Similar results also obtains for the logistic specification.

4. Model Selection Versus Hypothesis Testing

Hypothesis testing and model selection are different strands in the model evaluation literature. However, these strands differ in a number of important respects which are worth emphasising here.[10] Model selection begins with a given set of models, $\mathcal{M}$, characterised by the (possibly) conditional p.d.fs

$$\mathcal{M}=\{f_i(\mathbf{Y}|\mathbf{X}_i,\boldsymbol{\psi}_i),i=1,2,...,m\},$$

with the aim of *choosing* one of the models under consideration for a particular purpose with a specific loss (utility) function in mind. In essence model selection is a part of decision making and as argued in Granger and Pesaran (1999) ideally it should be fully integrated into the decision making process. However, most of the current literature on model selection builds on statistical measures of fit such as sums of squares of residuals or more generally maximized log-likelihood values, rather than economic value which one would expect to follow from a model choice.[11] As a result model selection seems much closer to hypothesis testing than it actually is in principle.

The model selection process treats all models under consideration symmetrically, while hypothesis testing attributes a different status to the null and to the alternative hypotheses and by design treats the models asymmetrically. Model selection always ends in a definite outcome, namely one of the models under consideration is selected for use in decision making. Hypothesis testing on the other hand asks whether there is any statistically significant evidence (in the Neyman-Pearson sense) of departure

[10] A review of the model selection literature is beyond the scope of the present paper. See, for example, Leamer (1983) for an excellent review. A recent review focussing upon the selection of regressors problems is to be found in Lavergne (1998). Two excellent texts are Grasa (1989) and Linhart and Zucchini (1986). Maddala (1981) edited a special issue of the Journal of Econometrics which focusses on model selection.

[11] For the case of the classical linear regression model examples of model selection criteria include Theil's $\bar{R}^2$, with more general loss functions based upon information criteria including Akaike's (Akaike (1973)) information criteria and Schwarz's (Schwarz (1978)) Bayesian information criterion.

from the null hypothesis in the direction of one or more alternative hypotheses. Rejection of the null hypothesis does not necessarily imply acceptance of any one of the alternative hypotheses; it only warns the investigator of possible shortcomings of the null that is being advocated. Hypothesis testing does not seek a definite outcome and if carried out with due care need not lead to a favourite model. For example, in the case of non-nested hypothesis testing it is possible for all models under consideration to be rejected, or all models to be deemed as observationally equivalent.

Due to its asymmetric treatment of the available models, the choice of the null hypothesis plays a critical role in the hypothesis testing approach. When the models are nested the most parsimonious model can be used as the null hypothesis. But in the case of non-nested models (particularly when the models are globally non-nested) there is no natural null, and it is important that the null hypothesis is selected on a priori grounds.[12] Alternatively, the analysis could be carried out with different models in the set treated as the null. Therefore, the results of non-nested hypothesis testing is less clear cut as compared to the case where the models are nested. [13].

It is also important to emphasise the distinction between *paired* and joint non-nested hypothesis tests. Letting f_1 denote the null model and $f_i \in \mathcal{M}$, $i = 2, ..., m$ index a set of $m-1$ alternative models, a paired test is a test of f_1 against a *single* member of $\mathcal{M}$, whereas a joint test is a test of f_1 against multiple alternatives in $\mathcal{M}$. McAleer (1995) is careful to highlight this distinction and in doing so points out a deficiency in many applied studies insofar as many authors have utilised a sequence of paired tests for problems characterised by multiple alternatives. Examples of studies which have applied non-nested tests to the choice between more than two models include Sawyer (1984), Smith and Maddala (1983) and Davidson and MacKinnon (1981). The paper by Sawyer is particularly relevant since he develops the multiple model equivalent of the Cox test.

The distinction between model selection and non-nested hypothesis tests can also be motivated from the perspective of Bayesian versus sampling-theory approaches to the problem of inference. For example, it is likely that with a large amount of data the posterior probabilities associated with a particular hypothesis will be close to one. However, the distinction drawn by Zellner (1971) between "comparing" and "testing" hypothesis is relevant given that within a Bayesian perspective the progression from a set of prior to posterior probabilities on $\mathcal{M}$, mediated by the Bayes factor, does not necessarily involve a decision to accept or reject the hypothesis. If a decision is required it is generally based upon minimising a particular expected loss function. Thus, model selection motivated by a decision problem is much more readily reconcilable with the Bayesian rather than the classical approach to model selection.

Finally, the choice between hypothesis testing and model selection clearly de-

[12] The concepts of globally and partially non-nested models are defined in Pesaran (1987).

[13] See also Dastoor (1981) for further discussion.

pends on the primary objective of the exercise. There are no definite rules. Model selection is more appropriate when the objective is decision making. Hypothesis testing is better suited to inferential problems where the empirical validity of a theoretical prediction is the primary objective. A model may be empirically adequate for a particular purpose but of little relevance for another use. Only in the unlikely event that the true model is known or knowable will the selected model be universally applicable. In the real world where the truth is elusive and unknowable both approaches to model evaluation are worth pursuing.

5. Alternative Approaches to Testing Non-nested Hypotheses with Application to Linear Regression Models

To provide an intuitive introduction to concepts which are integral to an understanding of non-nested hypothesis tests we consider testing of linear regression models as a convenient starting point. In the ensuing discussion we demonstrate that despite its special features non-nested hypothesis testing is firmly rooted within the Neyman-Pearson framework.

There are three general approaches to non-nested hypothesis testing all discussed in the pioneering contributions of Cox (1961) and Cox (1962). (i)- The modified (centred) log-likelihood ratio procedure, also known as the Cox test. (ii)- The comprehensive model approach, whereby the non-nested models are tested against an artificially constructed general model that includes the non-nested models as special cases. This approach was advocated by Atkinson (1970) and was later taken up under a different guise by Davidson and MacKinnon (1981) in developing their J-test and by Fisher and McAleer (1981) who proposed a related alternative procedure known as the JA-test. (iii)- A third approach, originally considered by Deaton (1982) and Dastoor (1983) and further developed by Gourieroux, Monfort, and Trognon (1983) and Mizon and Richard (1986) is the encompassing procedure where ability of one model to explain particular features of an alternative model is tested directly. The Wald and Score Encompassing Tests (usually denoted by WET and SET) are typically constructed under the assumption that one of the rival models is correct. Encompassing tests when the true model does not necessarily lie in the set of models (whether nested or non-nested) under consideration are proposed by Gourieroux and Monfort (1995) and Smith (1993).

We shall now illustrate the main features of these three approaches in the context of the classical linear normal regression models (3.1) and (3.2) set out above. Rewriting these models in familiar matrix notations we have:

$$H_f \quad : \quad \mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{u}_{f,} \qquad \mathbf{u}_f \sim N(\mathbf{0}, \sigma^2\mathbf{I}_T), \tag{5.1}$$

$$H_g \quad : \quad \mathbf{y} = \mathbf{Z}\boldsymbol{\beta} + \mathbf{u}_g, \qquad \mathbf{u}_g \sim N(\mathbf{0}, \omega^2\mathbf{I}_T), \tag{5.2}$$

where $\mathbf{y}$ is the $T \times 1$ vector of observations on the dependent variable, $\mathbf{X}$ and $\mathbf{Z}$ are

$T \times k_f$ and $T \times k_g$ observation matrices for the regressors of models H_f and H_g, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the $k_f \times 1$ and $k_g \times 1$ unknown regression coefficient vectors, $\mathbf{u}_f$ and $\mathbf{u}_g$ are the $T \times 1$ disturbance vectors, and $\mathbf{I}_T$ is an identity matrix of order T. In addition, throughout this section we assume that

$$T^{-1}\mathbf{X}'\mathbf{u}_f \xrightarrow{p} \mathbf{0},\ T^{-1}\mathbf{X}'\mathbf{u}_g \xrightarrow{p} \mathbf{0}, T^{-1/2}\mathbf{X}'\mathbf{u}_f \overset{a}{\sim} N(\mathbf{0}, \sigma^2\Sigma_{xx}),$$
$$T^{-1}\mathbf{Z}'\mathbf{u}_g \xrightarrow{p} \mathbf{0},\ T^{-1}\mathbf{Z}'\mathbf{u}_f \xrightarrow{p} \mathbf{0},\ T^{-1/2}\mathbf{Z}'\mathbf{u}_g \overset{a}{\sim} N(\mathbf{0}, \omega^2\Sigma_{zz}),$$

$$\hat{\Sigma}_{xx} = T^{-1}\mathbf{X}'\mathbf{X} \xrightarrow{p} \Sigma_{xx},\ \hat{\Sigma}_{zz} = T^{-1}\mathbf{Z}'\mathbf{Z} \xrightarrow{p} \Sigma_{zz},\ \hat{\Sigma}_{zx} = T^{-1}\mathbf{Z}'\mathbf{X} \xrightarrow{p} \Sigma_{zx},$$

where $\xrightarrow{p}$ denotes convergence in probability, the matrices $\hat{\Sigma}_{xx}$, Σ_{xx}, $\hat{\Sigma}_{zz}, \Sigma_{zz}$ are non-singular, $\Sigma_{zx} = \Sigma_{xz}' \neq \mathbf{0}$, and set

$$\Sigma_f = \Sigma_{xx} - \Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{zx}, \text{ and } \Sigma_g = \Sigma_{zz} - \Sigma_{zx}\Sigma_{xx}^{-1}\Sigma_{xz}.$$

5.1. Motivation for Non-Nested Statistics

From a statistical view point the main difference between the nested and non-nested hypothesis testing lies in the fact that the usual log-likelihood ratio or Wald statistics used in the conventional hypothesis testing are automatically centred at zero under the null when the hypotheses under consideration are nested while this is not true in the case of non-nested hypotheses. However, once the conventional test statistics are appropriately centred (at least asymptotically) the same classical techniques can be applied to testing of non-nested hypotheses. Using the two non-nested linear regression models in (5.1) and (5.2) we first demonstrate the problems with standard test statistics by focussing on a simple comparison of sums of squared errors.

Consider the following test statistic:

$$\xi_T = \tilde{\sigma}_g^2 - \tilde{\sigma}_f^2 \tag{5.3}$$

where

$$\begin{aligned} \tilde{\sigma}_f^2 &= \mathbf{e}_f'\mathbf{e}_f/(T - k_f) \\ \tilde{\sigma}_g^2 &= \mathbf{e}_g'\mathbf{e}_g/(T - k_g) \end{aligned}$$

and $\mathbf{e}_f$ is the OLS residual vector under H_f such that $\mathbf{e}_f = \mathbf{M}_f\mathbf{y}$. Note that (5.3) represents a natural starting point being the difference between the mean sum of squared errors for the two models.

In general the exact distribution of ξ_T will depend on the unknown parameters. To see this first note that under $H_f, \mathbf{e}_f = \mathbf{M}_f(\mathbf{u}_f + \mathbf{X}\boldsymbol{\alpha})$ therefore, (since $\mathbf{M}_f\mathbf{X} = 0$), we have

$$(T - k_f)\tilde{\sigma}_f^2 = \mathbf{u}_f'\mathbf{M}_f\mathbf{u}_f. \tag{5.4}$$

Now under H_f,

$$\mathbf{e}_g = \mathbf{M}_g\mathbf{y} = \mathbf{M}_g(\mathbf{X}\boldsymbol{\alpha} + \mathbf{u}_f),$$

or

$$\mathbf{e}_g = \mathbf{M}_g\mathbf{X}\boldsymbol{\alpha} + \mathbf{M}_g\mathbf{u}_f$$

and

$$\begin{aligned}(T-k_g)\tilde{\sigma}_g^2 &= \mathbf{e}_g'\mathbf{e}_g \\ &= (\mathbf{u}_f' + \boldsymbol{\alpha}'\mathbf{X}')\mathbf{M}_g(\mathbf{X}\boldsymbol{\alpha} + \mathbf{u}_f) \\ &= \mathbf{u}_f'\mathbf{M}_g\mathbf{u}_f + 2\boldsymbol{\alpha}'\mathbf{X}'\mathbf{M}_g\mathbf{u}_f + \boldsymbol{\alpha}'\mathbf{X}'\mathbf{M}_g\mathbf{X}\boldsymbol{\alpha}.\end{aligned} \tag{5.5}$$

Using (5.4) and (5.5) in (5.3) and taking expectations (under H_f) we have

$$E(\xi_T) = \frac{\boldsymbol{\alpha}'\mathbf{X}'\mathbf{M}_g\mathbf{X}\boldsymbol{\alpha}}{T-k_g} \geq 0 \tag{5.6}$$

which we denote by $\mu_T = (\boldsymbol{\alpha}'\mathbf{X}'\mathbf{M}_g\mathbf{X}\alpha)/(T-k_g)$. Since ξ_T does not have mean zero under the null hypothesis H_f, then ξ_T cannot provide us with a suitable *test-statistic*. Notice, however that when H_f is nested within H_g, then $\mathbf{M}_g\mathbf{X} = 0$ and ξ_T will have mean zero (exactly) under H_g. In this case if we also assume that $\mathbf{u}_f$ is normally distributed it can be easily shown that

$$\frac{(T-k_f)\xi_T}{r\bar{\sigma}_g^2} = 1 - F_{r,T-k_g}$$

where $F_{r,T-k_g}$ is distributed as a (central) F with r and $T-k_g$ degrees of freedom; r here stands for the number of restrictions that we need to impose on H_g in order to obtain H_f.

A fundamental tenet of classical hypothesis testing is that the distribution of the test statistic is known under a well specified null hypothesis. Thus, in this context if H_f is nested within H_g then under the null of H_f the normalised difference between the sum of squared errors has a zero expectation. When H_f is not nested within H_g we may adopt a number of alternate approaches. First, a suitable test statistic that has zero mean asymptotically will be

$$z_T = \xi_T - \hat{\mu}_T$$

where $\hat{\mu}_T$ is a consistent estimator of μ_T under H_f. More specifically

$$z_T = \tilde{\sigma}_g^2 - \tilde{\sigma}_f^2 - \frac{\hat{\boldsymbol{\alpha}}'\mathbf{X}'\mathbf{M}_g\mathbf{X}\hat{\boldsymbol{\alpha}}}{T-k_g}, \tag{5.7}$$

where $\hat{\boldsymbol{\alpha}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. (5.7) represents an example of *centring* of a test statistic such that the distribution of z_T is known (asymptotically). Cox (1961), Cox (1962)

utilised this approach to centre the log-likelihood ratio statistic for two non-nested models. When the models are nested the log-likelihood ratio statistic is properly centred (at least asymptotically). For example, if we let $L_f(\boldsymbol{\theta})$ and $L_g(\boldsymbol{\gamma})$ denote, respectively the log-likelihood functions for H_f and H_g, and we if assume that H_f is nested within H_g, then under H_f the log-likelihood ratio statistic, $2\left[L_g(\hat{\boldsymbol{\gamma}}_T) - L_f(\hat{\boldsymbol{\theta}}_T)\right]$, does not require any centring and the test defined by the critical region

$$2\left[L_g(\hat{\boldsymbol{\gamma}}_T) - L_f(\hat{\boldsymbol{\theta}}_T)\right] \geq \chi^2_{(1-\alpha)}(r)$$

where r is the number of parameter restrictions required to obtain H_f from H_g, asymptotically has the size α and is consistent. In the case of non-nested models the likelihood ratio statistic is not distributed as a chi-squared random variable. The reason for this is simple. The degrees of freedom of the chi-square statistic for the LR test is equal to the reduction in the size of the parameter space after imposing the necessary set of zero restrictions. Thus, if neither H_f nor H_g nests the other model, the attendant parameter spaces and hence the likelihoods are unrelated. In Section 5.2 we examine the application of centring (or mean adjustment) of the likelihood ratio statistic to obtain a test statistic that has a known asymptotic distribution. Given that in most instances the form of mean adjustment involves analytically intractable expectations in Section 7.1 we examine the use of simulation methods as a method of circumventing this problem.

Following seminal work by Efron (1979), an alternative approach conducts inference utilising the empirical distribution function of the test statistic. In this instance there is, in general, no need to centre ξ_T using $\hat{\mu}_T$. Instead we take ξ_T as the observed test statistic, and given a null hypothesis, we simulate a large number, say R, of the $\tilde{\sigma}^2_g$, $\tilde{\sigma}^2_f$ pairs. The empirical distribution function for ξ_T is then constructed based on $\hat{\sigma}^2_{gr}$ and $\hat{\sigma}^2_{fr}$, $r = 1, 2, ..., R$. In Section 7.2 we examine the use of bootstrap procedures for conducting non-nested hypothesis tests. We also consider the case for combining the type of mean adjustment in (5.7) with bootstrap procedures.

5.2. The Cox Procedure

This procedure focuses on the log-likelihood ratio statistic, and in the case of the above regression models is given by (using the notations of Section 2)

$$LR_{fg} = L_f(\hat{\boldsymbol{\theta}}_T) - L_g(\hat{\boldsymbol{\gamma}}_T) = \frac{T}{2}\ln\left(\frac{\hat{\sigma}^2_T}{\hat{\omega}^2_T}\right),$$

where

$$\begin{aligned} \hat{\sigma}^2_T &= T^{-1}\mathbf{e}_f'\mathbf{e}_f,\ \hat{\boldsymbol{\alpha}}_T = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \\ \mathbf{e}_f &= \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\alpha}}_T = \mathbf{M}_x\mathbf{y},\ \mathbf{M}_x = \mathbf{I}_T - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}', \end{aligned} \tag{5.8}$$

and

$$\begin{aligned}\hat{\omega}_T^2 &= T^{-1}\mathbf{e}_g'\mathbf{e}_g,\ \hat{\boldsymbol{\beta}}_T = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}, \\ \mathbf{e}_g &= \mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\beta}}_T = \mathbf{M}_z\mathbf{y},\ \mathbf{M}_z = \mathbf{I}_T - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'. \end{aligned} \tag{5.9}$$

In the general case where the regression models are non-nested the average log-likelihood ratio statistic, $\frac{1}{2}\ln\left(\hat{\sigma}_T^2/\hat{\omega}_T^2\right)$, does not converge to zero even if T is sufficiently large. For example, under H_f we have

$$P\lim_{T\to\infty}\left(T^{-1}LR_{fg} \mid H_f\right) = \frac{1}{2}\ln\left(\frac{\sigma_0^2}{\omega_*^2(\boldsymbol{\theta}_0)}\right) = \frac{1}{2}\ln\left(\frac{\sigma_0^2}{\sigma_0^2 + \boldsymbol{\alpha}_0'\Sigma_f\boldsymbol{\alpha}_0}\right),$$

and under H_g:

$$P\lim_{T\to\infty}\left(T^{-1}LR_{fg} \mid H_g\right) = \frac{1}{2}\ln\left(\frac{\sigma_*^2(\boldsymbol{\gamma}_0)}{\omega_0^2}\right) = \frac{1}{2}\ln\left(\frac{\omega_0^2 + \boldsymbol{\beta}_0'\Sigma_g\boldsymbol{\beta}_0}{\omega_0^2}\right).$$

The LR statistic is naturally centred at zero if one or the other of the above probability limits is equal to zero; namely if either $\Sigma_f = \mathbf{0}$ or $\Sigma_g = \mathbf{0}$.[14] When $\Sigma_f = \mathbf{0}$ then $\mathbf{X} \subset \mathbf{Z}$ and H_f is nested in H_g. Alternatively, if $\Sigma_g = \mathbf{0}$, then $\mathbf{Z} \subset \mathbf{X}$ and H_g is nested in H_f. Finally, if both $\Sigma_f = \mathbf{0}$ and $\Sigma_g = \mathbf{0}$ then the two regression models are observationally equivalent. In the non-nested case where both $\Sigma_f \neq \mathbf{0}$ and $\Sigma_g \neq \mathbf{0}$, the standard LR statistic will not be applicable and needs to be properly centred. Cox's contribution was to note that this problem can be overcome if a consistent estimate of $P\lim_{T\to\infty}\left(T^{-1}LR_{fg} \mid H_f\right)$, which we denote by $\hat{E}_f(T^{-1}LR_{fg})$, is subtracted from $T^{-1}LR_{fg}$, which yields the new centred (modified) log-likelihood ratio statistic (also known as the Cox statistic) for testing H_f against H_g:

$$\begin{aligned} S_{fg} &= T^{-1}LR_{fg} - \hat{E}_f(T^{-1}LR_{fg}) \qquad (5.10) \\ &= \frac{1}{2}\ln\left(\frac{\hat{\sigma}_T^2}{\hat{\omega}_T^2}\right) - \frac{1}{2}\ln\left(\frac{\sigma_T^2}{\hat{\sigma}_T^2 + \hat{\boldsymbol{\alpha}}_T'\hat{\Sigma}_f\hat{\boldsymbol{\alpha}}_T}\right) \\ &= \frac{1}{2}\ln\left(\frac{\hat{\sigma}_T^2 + \hat{\boldsymbol{\alpha}}_T'\hat{\Sigma}_f\hat{\boldsymbol{\alpha}}_T}{\hat{\omega}_T^2}\right). \qquad (5.11) \end{aligned}$$

It is now clear that by construction the Cox statistic, S_{fg}, has asymptotically mean zero under H_f. As was pointed out earlier, since there is no natural null hypothesis in this set up, one also needs to consider the modified log-likelihood ratio statistic for testing H_g against H_f which is given by

$$S_{gf} = \frac{1}{2}\ln\left(\frac{\hat{\omega}_T^2 + \hat{\boldsymbol{\beta}}_T'\hat{\Sigma}_g\hat{\boldsymbol{\beta}}_T}{\hat{\sigma}_T^2}\right).$$

[14] The cases where $\Sigma_f \neq \mathbf{0}$ (respectively $\Sigma_g \neq \mathbf{0}$) but neverthelss $\Sigma_f\boldsymbol{\alpha}_0 = 0$ (respectively $\Sigma_g\boldsymbol{\beta}_0 = 0$) are discussed in Pesaran (1987, p. 74).

Both of these test statistics (when appropriately normalized by $\sqrt{T}$) are asymptotically normally distributed under their respective nulls with a zero mean and finite variances. For the test of H_f against H_g we have[15]

$$\widehat{Asyvar}(\sqrt{T}S_{fg}) = V_{fg} = \frac{\hat{\sigma}_T^2\left(\hat{\boldsymbol{\alpha}}_T'\mathbf{X}'\mathbf{M}_z\mathbf{M}_x\mathbf{M}_z\mathbf{X}\hat{\boldsymbol{\alpha}}_T\right)}{T\left(\hat{\sigma}_T^2 + \hat{\boldsymbol{\alpha}}_T'\hat{\Sigma}_f\hat{\boldsymbol{\alpha}}_T\right)^2}.$$

The associated standardized Cox statistic is given by

$$N_{fg} = \frac{\sqrt{T}S_{fg}}{\sqrt{V_{fg}}} \overset{a}{\backsim} N(0,1). \tag{5.12}$$

By reversing the role of the null and the alternative hypothesis a similar standardized Cox statistic can be computed for testing H_g against H_f, which we denote by N_{gf}. Denote the $(1-\alpha)$ percent critical value of the standard normal distribution by C_α, then four outcomes are possible:

(1) Reject H_g but not H_f if $\mid N_{fg} \mid < C_\alpha$ and $\mid N_{gf} \mid \geq C_\alpha$,
(2) Reject H_f but not H_g if $\mid N_{fg} \mid \geq C_\alpha$ and $\mid N_{gf} \mid < C_\alpha$,
(3) Reject both H_f and H_g if $\mid N_{fg} \mid \geq C_\alpha$ and $\mid N_{gf} \mid \geq C_\alpha$,
(4) Reject neither H_f or H_g if $\mid N_{fg} \mid < C_\alpha$ and $\mid N_{gf} \mid < C_\alpha$.

These are to be contrasted to the outcomes of the nested hypothesis testing where the null is either rejected or not, which stem from the fact that when the hypotheses under consideration are non-nested there is no natural null (or maintained) hypothesis and one therefore needs to consider in turn each of the hypotheses as the null. So there are twice as many possibilities as there are when the hypotheses are nested. Note that if we utilise the information in the *direction of rejection,* that is instead of comparing the *absolute* value of N_{fg} with C_α we determine whether rejection is in the direction of the null or the alternative, there are a total of eight possible test outcomes (see the discussion in Fisher and McAleer (1979) and Dastoor (1981)). This aspect of non-nested hypothesis testing has been criticized by some commentators; pointing out the test outcome can lead to ambiguities. (See, for example, Granger, King, and White (1995)). However, this is a valid criticism only if the primary objective is to *select* a specific model for forecasting or decision making, but not if the aim is to learn about the comparative strengths and weaknesses of rival explanations. What is viewed as a weakness from the perspective of model selection now becomes a strength when placed in the context of statistical inference and model building. For example, when both models are rejected the analysis points the investigator in the direction of developing a third model which incorporates the main desirable features of the original, as well as being theoretically meaningful. (See Pesaran and Deaton (1978)).

[15] See Pesaran (1974) for details of the derivations.

5.3. The Comprehensive Approach

Another approach closely related to the Cox's procedure is the comprehensive approach advocated by Atkinson (1970) whereby tests of non-nested models are based upon a third comprehensive model, artificially constructed so that each of the non-nested models can be obtained from it as special cases. Clearly, there are a large number of ways that such a comprehensive model can be constructed. A prominent example is the exponential mixture, H_λ, which in the case of the non-nested models (2.4) and (2.5) is defined by

$$H_\lambda: \quad c_\lambda(\mathbf{y}_t|\mathbf{x}_t,\mathbf{z}_t,\Omega_{t-1};\boldsymbol{\theta},\boldsymbol{\gamma}) = \frac{f(\mathbf{y}_t|\mathbf{x}_t,\Omega_{t-1};\boldsymbol{\theta})^{1-\lambda}g(\mathbf{y}_t|\mathbf{z}_t,\Omega_{t-1};\boldsymbol{\gamma})^{\lambda}}{\int_{\mathcal{R}_y} f(\mathbf{y}_t|\mathbf{x}_t,\Omega_{t-1};\boldsymbol{\theta})^{1-\lambda}g(\mathbf{y}_t|\mathbf{z}_t,\Omega_{t-1};\boldsymbol{\gamma})^{\lambda}d\mathbf{y}_t},$$

where $\mathcal{R}_y$ represents the domain of variations of $\mathbf{y}_t$, and the integral in the denominator ensures that the combined function, $c_\lambda(\mathbf{y}_t|\mathbf{x}_t,\mathbf{z}_t,\Omega_{t-1};\boldsymbol{\theta},\boldsymbol{\gamma})$, is in fact a proper density function integrating to unity over $\mathcal{R}_y$. The "mixing" parameter λ varies in the range $[0,1]$ and represents the weight attached to model H_f. A test of $\lambda = 0$ ($\lambda = 1$) against the alternative that $\lambda \neq 0$ ($\lambda \neq 1$) can now be carried out using standard techniques from the literature on nested hypothesis testing. (See Atkinson (1970) and Pesaran (1982a)). This approach is, however, subject to three important limitations. First, although the testing framework is nested, the test of $\lambda = 0$ is still *non-standard* due to the fact that under $\lambda = 0$ the parameters of the alternative hypothesis, $\boldsymbol{\gamma}$, disappears. This is known as the Davies's problem. (See Davies (1977)). The same also applies if the interest is in testing $\lambda = 1$. The second limitation is due to the fact that testing $\lambda = 0$ against $\lambda \neq 0$, is not equivalent to testing H_f against H_g, which is the problem of primary interest. This implicit change of the alternative hypothesis can have unfavourable consequences for the power of non-nested tests. Finally, the particular functional form used to combine the two models is arbitrary and does not allow identification of the mixing parameter, λ, even if $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ are separately identified under H_f and H_g respectively. (See Pesaran (1981)).

The application of the comprehensive approach to the linear regression models (5.1) and (5.2) yields:

$$H_\lambda: \quad \mathbf{y} = \left\{\frac{(1-\lambda)\nu^2}{\sigma^2}\right\}\mathbf{X}\boldsymbol{\alpha} + \left\{\frac{\lambda\nu^2}{\omega^2}\right\}\mathbf{Z}\boldsymbol{\beta} + \mathbf{u}, \quad \mathbf{u} \sim N(\mathbf{0},\nu^2\mathbf{I}_T), \tag{5.13}$$

where $\nu^{-2} = (1-\lambda)\sigma^{-2} + \lambda\omega^{-2}$. It is clear that the mixing parameter λ is not identified.[16] In fact setting $\kappa = \lambda\nu^2/\omega^2$ the above "combined" regression can also be written as

$$H_\kappa: \quad \mathbf{y} = (1-\kappa)\mathbf{X}\boldsymbol{\alpha} + \kappa\mathbf{Z}\boldsymbol{\beta} + \mathbf{u}, \tag{5.14}$$

[16]For example, it is not possible to test whether $\lambda = 1/2$, which could have been of interest in assessing the relative weights attached to the two rival models.

and a test of $\lambda = 0$ in (5.13) can be carried by testing $\kappa = 0$ in (5.14). Since the error variances σ^2 and ω^2 are strictly positive $\lambda = 0$ will be equivalent to testing $\kappa = 0$. The Davies problem, of course, continues to apply and under H_f ($\kappa = 0$) the coefficients of the rival model, $\boldsymbol{\beta}$, disappear from the combined model. To resolve this problem Davies (1977) proposes a two-stage procedure. First, for a given value of $\boldsymbol{\beta}$ a statistic for testing $\kappa = 0$ is chosen. In the present application this is given by the t-ratio of κ in the regression of $\mathbf{y}$ on $\mathbf{X}$ and $\mathbf{y}_\beta = \mathbf{Z}\boldsymbol{\beta}$, namely

$$t_\kappa(\mathbf{Z}\boldsymbol{\beta}) = \frac{\boldsymbol{\beta}'\mathbf{Z}'\mathbf{M}_x\mathbf{y}}{\hat{\nu}\left(\boldsymbol{\beta}'\mathbf{Z}'\mathbf{M}_x\mathbf{Z}\boldsymbol{\beta}\right)^{1/2}},$$

$$\hat{\nu}^2 = \frac{1}{T - k_f - 1}\left\{\mathbf{y}'\mathbf{M}_x\mathbf{y} - \frac{\left(\boldsymbol{\beta}'\mathbf{Z}'\mathbf{M}_x\mathbf{y}\right)^2}{\boldsymbol{\beta}'\mathbf{Z}'\mathbf{M}_x\mathbf{Z}\boldsymbol{\beta}}\right\},$$

and where $\mathbf{M}_x$ is already defined by (5.8). In the second stage a test is constructed based on the entire random function of $t_\kappa(\mathbf{Z}\boldsymbol{\beta})$ viewed as a function of $\boldsymbol{\beta}$. One possibility would be to construct a test statistic based on

$$F_\kappa = \underset{\boldsymbol{\beta}}{Max}\left\{t_\kappa(\mathbf{Z}\boldsymbol{\beta})\right\}.$$

Alternatively, a test statistic could be based on the average value of $t_\kappa(\mathbf{Z}\boldsymbol{\beta})$ obtained using a suitable prior distribution for $\boldsymbol{\beta}$. Following the former classical route it is then easily seen that F_κ becomes the standard F_{z^*} statistic for testing $\mathbf{b}_2 = \mathbf{0}$, in the regression

$$\mathbf{y} = \mathbf{X}\mathbf{b}_1 + \mathbf{Z}^*\mathbf{b}_2 + \mathbf{v}_f, \tag{5.15}$$

where $\mathbf{Z}^*$ is the set of regressors in $\mathbf{Z}$ but not in $\mathbf{X}$, namely $\mathbf{Z}^* = \mathbf{Z} - \mathbf{X} \cap \mathbf{Z}$.[17] Similarly for testing H_g against H_f the comprehensive approach involves testing $\mathbf{c}_1 = \mathbf{0}$, in the combined regression

$$\mathbf{y} = \mathbf{X}^*\mathbf{c}_1 + \mathbf{Z}\mathbf{c}_2 + \mathbf{v}_g, \tag{5.16}$$

where $\mathbf{X}^*$ is the set of variables in $\mathbf{X}$ but not in $\mathbf{Z}$. Denoting the F statistic for testing $\mathbf{c}_1 = \mathbf{0}$ in this regression by F_{x^*} , notice that there are still four possible outcomes to this procedure; in line with the ones detailed above for the Cox test. This is because we have two F statistics, F_{x^*} and F_{z^*}, with the possibility of rejecting both hypotheses, rejecting neither, etc.

An altogether different approach to the resolution of the Davies' problem would be to replace the regression coefficients, $\boldsymbol{\beta}$, in (5.14) by an estimate, say $\tilde{\boldsymbol{\beta}}$, and then proceed as if $\tilde{\mathbf{y}}_\beta = \mathbf{Z}\tilde{\boldsymbol{\beta}}$ is data. This is in effect what is proposed by Davidson and MacKinnon (1981) and Fisher and McAleer (1981). Davidson and MacKinnon

[17] For a proof see McAleer and Pesaran (1986).

suggest using the estimate of $\boldsymbol{\beta}$ under H_g, namely $\hat{\boldsymbol{\beta}}_T = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}\mathbf{y}$. This leads to the J test which is the standard t ratio of the estimate of κ in the artificial regression[18]

$$H_\kappa : \quad \mathbf{y} = \mathbf{X}\mathbf{a} + \kappa\mathbf{Z}\hat{\boldsymbol{\beta}}_T + \mathbf{v}_\kappa. \tag{5.17}$$

For testing H_g against H_f, the J test will be based on the OLS regression of $\mathbf{y}$ on $\mathbf{Z}$ and $\mathbf{X}\hat{\boldsymbol{\alpha}}_T$, and the J statistic is the t-ratio of the coefficient of $\mathbf{X}\hat{\boldsymbol{\alpha}}_T$ (which is the vector of fitted values under H_f) in this regression.

The test proposed by Fisher and McAleer (known as the JA test) replaces $\boldsymbol{\beta}$ by the estimate of its pseudo-true value under H_f, given by $\boldsymbol{\beta}_*(\hat{\boldsymbol{\alpha}}_T)$

$$\hat{\boldsymbol{\beta}}_*(\hat{\boldsymbol{\alpha}}_T) = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\hat{\boldsymbol{\alpha}}_T.$$

In short the JA test of H_f against H_g is the t-ratio of the coefficient of $\hat{\mathbf{y}}_{\beta\alpha} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\hat{\boldsymbol{\alpha}}_T$ in the OLS regression of $\mathbf{y}$ on $\mathbf{X}$ and $\hat{\mathbf{y}}_{\beta\alpha}$. Similarly, a JA test of H_g against H_f can be computed.

Both the J and the JA test statistics, as well as their various variations proposed in the literature can also be derived as linear approximations to the Cox test statistic. See (5.10).

Various extensions of the non-nested hypothesis testing have also appeared in the literature. These include tests of non-nested linear regression models with serially correlated errors (McAleer, Pesaran, and Bera (1990)), models estimated by instrumental variables (Ericsson (1983) and Godfrey (1983)), models estimated by the generalized method of moments (Smith (1992)), non-nested Euler equations (Ghysels and Hall (1990)), autoregressive versus moving average models (Walker (1967), King (1983)), generalized autoregressive conditional heteroscedastic (GARCH) model against the exponential-GARCH model (McAleer and Ling (1998)), linear versus log-linear models (Aneuryn-Evans and Deaton (1980), Davidson and MacKinnon (1985), Pesaran and Pesaran (1995)), logit and probit models (Pesaran and Pesaran (1993), Weeks (1996) and Duncan and Weeks (1998)), non-nested threshold autoregressive models (Altissimo and Violante (1998), Pesaran and Potter (1997) and Kapetanios and Weeks (1999)).

5.4. The Encompassing Approach

This approach generalizes the Cox's original idea and asks whether model H_f can explain one or more features of the rival model H_g. When *all* the features of model H_g can be explained by model H_f it is said that model H_f *encompasses* model H_g; likewise model H_g is said to encompass model H_f if all the features of model H_f can be explained by model H_g. A formal definition of encompassing can be given in terms of the pseudo-true parameters and the binding functions defined in Section 2:

[18] Chao and Swanson (1997) provide some asymptotic results for the J test in the case of non-nested models with $I(1)$ regressors.

Model H_g is said to encompass model H_f, respectively defined by (2.5) and (2.4), if and only if

$$H_g \mathcal{E} H_f : \quad \boldsymbol{\theta}_{h*} = \boldsymbol{\theta}_*(\boldsymbol{\gamma}_{h*}). \tag{5.18}$$

Similarly, H_f is said to encompass H_g (or H_g is encompassed by H_f) if and only if

$$H_f \mathcal{E} H_g : \quad \boldsymbol{\gamma}_{h*} = \boldsymbol{\gamma}_*(\boldsymbol{\theta}_{h*}).$$

Recall that $\boldsymbol{\theta}_{h*}$ and $\boldsymbol{\gamma}_{h*}$ are the pseudo-true values of $\boldsymbol{\theta}$, and $\boldsymbol{\gamma}$ with respect to the true model H_h, and $\boldsymbol{\theta}_*(\cdot)$ and $\boldsymbol{\gamma}_*(\cdot)$ are the binding functions linking the parameters of the models H_f and H_g. For example, in the case of the linear rival regression models (3.1) and (3.2), and assuming that the true model is given by (3.5) then it is easily seen that the functions that bind the parameters of model H_g to that of H_f are

$$\boldsymbol{\theta}_*(\boldsymbol{\gamma}_{h*}) = \begin{pmatrix} \hat{\Sigma}_{xx}^{-1}\hat{\Sigma}_{xz}\boldsymbol{\beta}_{h*} \\ \omega_{h*}^2 + \boldsymbol{\beta}_{h*}'(\hat{\Sigma}_{zz} - \hat{\Sigma}_{zx}\hat{\Sigma}_{xx}^{-1}\hat{\Sigma}_{xz})\boldsymbol{\beta}_{h*} \end{pmatrix}.$$

Using (3.7) to substitute for the pseudo-true values $\boldsymbol{\beta}_{h*}$ and ω_{h*}^2 we have

$$\boldsymbol{\theta}_*(\boldsymbol{\gamma}_{h*}) = \begin{pmatrix} \hat{\Sigma}_{xx}^{-1}\hat{\Sigma}_{xz}\hat{\Sigma}_{zz}^{-1}\hat{\Sigma}_{zw}\boldsymbol{\delta} \\ \upsilon^2 + \boldsymbol{\delta}'(\hat{\Sigma}_{ww} - \hat{\Sigma}_{wz}\hat{\Sigma}_{zz}^{-1}\hat{\Sigma}_{zw})\boldsymbol{\delta} + \boldsymbol{\delta}'\hat{\Sigma}_{wz}\hat{\Sigma}_{zz}^{-1}(\hat{\Sigma}_{zz} - \hat{\Sigma}_{zx}\hat{\Sigma}_{xx}^{-1}\hat{\Sigma}_{xz})\hat{\Sigma}_{zz}^{-1}\hat{\Sigma}_{zw}\boldsymbol{\delta} \end{pmatrix}.$$

Therefore, conditional on the observation matrices $\mathbf{X}$, $\mathbf{Z}$, and $\mathbf{W}$, model H_f encompasses model H_g if and only if

$$\begin{aligned} & \begin{pmatrix} \hat{\Sigma}_{xx}^{-1}\hat{\Sigma}_{xw}\boldsymbol{\delta} \\ \upsilon^2 + \boldsymbol{\delta}'(\hat{\Sigma}_{ww} - \hat{\Sigma}_{wx}\hat{\Sigma}_{xx}^{-1}\hat{\Sigma}_{xw})\boldsymbol{\delta} \end{pmatrix} \\ = & \begin{pmatrix} \hat{\Sigma}_{xx}^{-1}\hat{\Sigma}_{xz}\hat{\Sigma}_{zz}^{-1}\hat{\Sigma}_{zw}\boldsymbol{\delta} \\ \upsilon^2 + \boldsymbol{\delta}'(\hat{\Sigma}_{ww} - \hat{\Sigma}_{wz}\hat{\Sigma}_{zz}^{-1}\hat{\Sigma}_{zw})\boldsymbol{\delta} + \boldsymbol{\delta}'\hat{\Sigma}_{wz}\hat{\Sigma}_{zz}^{-1}(\hat{\Sigma}_{zz} - \hat{\Sigma}_{zx}\hat{\Sigma}_{xx}^{-1}\hat{\Sigma}_{xz})\hat{\Sigma}_{zz}^{-1}\hat{\Sigma}_{zw}\boldsymbol{\delta} \end{pmatrix}. \end{aligned}$$

These conditions are simplified to

$$\hat{\Sigma}_{xw}\boldsymbol{\delta} = \hat{\Sigma}_{xz}\hat{\Sigma}_{zz}^{-1}\hat{\Sigma}_{zw}\boldsymbol{\delta}, \tag{5.19}$$

and

$$\boldsymbol{\delta}'\hat{\Sigma}_{wx}\hat{\Sigma}_{xx}^{-1}\hat{\Sigma}_{xw}\boldsymbol{\delta} = \boldsymbol{\delta}'\hat{\Sigma}_{wz}\hat{\Sigma}_{zz}^{-1}\hat{\Sigma}_{zx}\hat{\Sigma}_{xx}^{-1}\hat{\Sigma}_{xz}\hat{\Sigma}_{zz}^{-1}\hat{\Sigma}_{zw}\boldsymbol{\delta}. \tag{5.20}$$

But it is easily verified that (5.19) implies (5.20), namely encompassing with respect to the regression coefficients imply encompassing with respect to the error variances. Therefore, H_f is encompassed by H_g if and only if $(\mathbf{X}'\mathbf{M}_z\mathbf{W})\boldsymbol{\delta} = \mathbf{0}$. This condition is clearly satisfied if either H_f is nested within H_g, $(\mathbf{X}'\mathbf{M}_z = \mathbf{0})$, or if H_g contains the true model, $(\mathbf{M}_z\mathbf{W} = \mathbf{0})$. The remaining possibility, namely when $(\mathbf{X}'\mathbf{M}_z\mathbf{W}) = \mathbf{0}$,

but the true value of $\boldsymbol{\delta}$, say $\boldsymbol{\delta}_0$, is such that $(\mathbf{X}'\mathbf{M}_z\mathbf{W})\boldsymbol{\delta}_0 = \mathbf{0}$, is a rather a low probability event.

The encompassing hypothesis, $H_g\mathcal{E}H_f$, (or $H_f\mathcal{E}H_g$) can now be tested using the encompassing statistics, $\sqrt{T}\left[\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_*(\hat{\boldsymbol{\gamma}}_T)\right]$, $\left(\text{ or } \sqrt{T}\left[\hat{\boldsymbol{\gamma}}_T - \boldsymbol{\gamma}_*(\hat{\boldsymbol{\theta}}_T)\right]\right)$. Gourieroux and Monfort (1995) show that under the encompassing hypothesis, $\boldsymbol{\theta}_{h*} = \boldsymbol{\theta}_*(\boldsymbol{\gamma}_{h*})$, and assuming certain requilarity conditions are met, $\sqrt{T}\left[\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_*(\hat{\boldsymbol{\gamma}}_T)\right]$ is asymptotically normally distributed with zero means and a variance covariance matrix that in general depends in a complicated way on the probability density functions of the rival models under consideration. Complications arise since because H_g need not belong to H_h. Two testing procedures are proposed, the Wald encompassing test (WET) and the score encompassing test (SET), both being difficult to implement. First, the binding functions $\boldsymbol{\theta}_*(\cdot)$ and $\boldsymbol{\gamma}_*(\cdot)$ are not always easy to derive. (But this problem also afflicts the implementation of the Cox procedure, see below). Second, and more importantly, the variance-covariance matrices of $\sqrt{T}\left[\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_*(\hat{\boldsymbol{\gamma}}_T)\right]$, $\left(\text{ or } \sqrt{T}\left[\hat{\boldsymbol{\gamma}}_T - \boldsymbol{\gamma}_*(\hat{\boldsymbol{\theta}}_T)\right]\right)$, are, in general, non-invertible and the construction of WET and SET statistics involve the use of generalized inverse and this in turn requires estimation of the rank of these covariance matrices. Alternative ways of dealing with these difficulties are considered in Gourieroux and Monfort (1995) and Smith (1993).

In the case of linear regression models full parameter encompassing (namely an encompassing exercise involving both regression coefficients and error variances) is unnecessary.[19] Focussing on regression coefficients the encompassing statistics for testing $H_g\mathcal{E}H_f$ is given by

$$\sqrt{T}\left[\hat{\boldsymbol{\alpha}}_T - \boldsymbol{\alpha}_*(\hat{\boldsymbol{\beta}}_T)\right] = \sqrt{T}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_z\mathbf{y}.$$

Under H_h, defined by (3.5),

$$\sqrt{T}\left[\hat{\boldsymbol{\alpha}}_T - \boldsymbol{\alpha}_*(\hat{\boldsymbol{\beta}}_T)\right] = \sqrt{T}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{M}_z\mathbf{W})\boldsymbol{\delta} + \sqrt{T}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_z\mathbf{u}_h,$$

where $\mathbf{u}_h \backsim N(\mathbf{0}, \upsilon^2\mathbf{I}_T)$.[20] Hence, under the encompassing hypothesis, $(\mathbf{X}'\mathbf{M}_z\mathbf{W})\boldsymbol{\delta} = \mathbf{0}$, the encompassing statistic $\sqrt{T}\left[\hat{\boldsymbol{\alpha}}_T - \boldsymbol{\alpha}_*(\hat{\boldsymbol{\beta}}_T)\right]$ is asymptotically normally distributed with mean zero and the covariance matrix $\upsilon^2\Sigma_{xx}^{-1}(\Sigma_{xx} - \Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{zx})\Sigma_{xx}^{-1}$. Therefore, the construction of a standardized encompassing test statistic requires a consistent estimate of υ^2, the error variance of the true regression model, and this does not seem possible without further assumptions about the nature of the true model. In the literature it is often (implicitly) assumed that the true model is contained in the union intersection of the rival models under consideration (namely $\mathbf{W} \equiv \mathbf{X} \cup \mathbf{Z}$)

[19] Recall that the encompassing condition (5.19) for the regression coefficients implies the condition (5.20) for error variance encompassing but not *vice versa*.

[20] Notice that the normality assumption is not needed and can be relaxed.

and υ^2 is then consistently estimated from a regression of $\mathbf{y}$ on $\mathbf{X} \cup \mathbf{Z}$. Under this additional assumption, the WET statistic for testing $H_g \mathcal{E} H_f$, is given by

$$\mathcal{E}_{gf} = \frac{\mathbf{y}'\mathbf{M}_z\mathbf{X}(\mathbf{X}'\mathbf{M}_z\mathbf{X})^{-}\mathbf{X}'\mathbf{M}_z\mathbf{y}}{\hat{\upsilon}^2},$$

where $\hat{\upsilon}^2$ is the estimate of the error variance of the regression of $\mathbf{y}$ on $\mathbf{X} \cup \mathbf{Z}$, and $(\mathbf{X}'\mathbf{M}_z\mathbf{X})^{-}$ is a generalised inverse of $\mathbf{X}'\mathbf{M}_z\mathbf{X}$. This matrix is rank deficient whenever $\mathbf{X}$ and $\mathbf{Z}$ have variables in common, namely if $\mathbf{X} \cap \mathbf{Z} = \mathbf{Q} \neq \mathbf{0}$. Let $\mathbf{X} = (\mathbf{X}_1, \mathbf{Q})$ and $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Q})$, then

$$\mathbf{X}'\mathbf{M}_z\mathbf{X} = \begin{pmatrix} \mathbf{X}_1'\mathbf{M}_z\mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

But it is easily seen that $\mathcal{E}_{gf}$ is invariant to the choice of the g-inverse used and is given by

$$\mathcal{E}_{gf} = \frac{\mathbf{y}'\mathbf{M}_z\mathbf{X}_1(\mathbf{X}_1'\mathbf{M}_z\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_z\mathbf{y}}{\hat{\upsilon}^2},$$

and is identical to the standard Wald statistic for testing the statistical significance of $\mathbf{X}_1$ in the OLS regression of $\mathbf{y}$ on $\mathbf{Z}$ and $\mathbf{X}_1$. This is perhaps not surprising, considering the (implicit) assumption concerning the true model being a union intersection of the rival regression models H_f and H_g.

Other encompassing tests can also be developed depending on the parameters of interest or their functions. For example, a *variance* encompassing test of $H_g \mathcal{E} H_f$ compares a consistent estimate of σ^2 with that of its pseudo-true value σ^2_{h*}, namely $\hat{\sigma}^2_T - \sigma^2_*(\hat{\boldsymbol{\gamma}}_T) = \hat{\sigma}^2_T - \left[\hat{\omega}^2_T + T^{-1}\hat{\boldsymbol{\beta}}'_T\mathbf{Z}'\mathbf{M}_x\mathbf{Z}\hat{\boldsymbol{\beta}}_T\right]$.[21] Under the encompassing hypothesis this statistic tends to zero, but its asymptotic distributions in general depends on H_h. In the case where H_g contains the true model the variance encompassing test will become asymptotically equivalent to the Cox and the J tests discussed above.

The encompassing approach can also be applied to the log-likelihood functions. For example, to test $H_g \mathcal{E} H_f$ one could use the encompassing log-likelihood ratio statistic $T^{-1}\{L_f(\hat{\boldsymbol{\theta}}_T) - L_f(\boldsymbol{\theta}_*(\hat{\boldsymbol{\gamma}}_T))\}$. This test can also be motivated using Cox's idea of centred log-likelihood ratio statistic, with the difference that the centring is now carried out under H_h rather than under H_g (or H_f). See Gourieroux and Monfort (1995) and Smith (1993) for details and difficulties involved in their implementation. Other relevant literature include Dastoor (1983), Gourieroux, Monfort, and Trognon (1983) and Mizon and Richard (1986).

5.5. Power and Finite Sample Properties

A number of studies have examined the small sample properties of non-nested tests. For a limited number of cases it is possible to determine the exact form of the test

[21] Similarly, the variance encompassing statistic for testing $H_f \mathcal{E} H_g$ is given by $\hat{\omega}^2_T - \left[\hat{\sigma}^2_T + T^{-1}\hat{\boldsymbol{\alpha}}'_T\mathbf{X}'\mathbf{M}_z\mathbf{X}\hat{\boldsymbol{\alpha}}'_T\right]$.

statistic and the sampling distribution. For example, Godfrey (1983) shows that under H_f if $\mathbf{X}$ and $\mathbf{Z}$ are non-stochastic with normal errors, then the JA-test has an exact $t(T - k_f - 1)$ distribution.[22] In the majority of cases the finite sample properties have been examined using Monte Carlo studies. A recurrent finding is that many Cox-type tests for non-nested regression models have a finite sample size which is significantly greater than the nominal level. Modifications based upon mean and variance adjustments have been proposed in Godfrey and Pesaran (1983), and are shown to affect a substantial improvement in finite sample performance. The authors demonstrate that in experimental designs allowing for non-nested models with either non-normal errors, different number of regressors, or a lagged dependent variable, the adjusted Cox-test performs favourably relative to the J test or F-test.[23] In the case of non-nested linear regression models, Davidson and McKinnon (1982) compared a number of variants of the Cox test with F, JA and J test.

An analysis of the power properties of non-tested tests has been undertaken using a number of approaches. In the case of nested models local alternatives are readily defined in terms of parameters that link the null to the alternative. Obviously in the case of models that are globally non-nested (i.e. the exponential and log-normal) this procedure is not possible. In the case of regression models Pesaran (1982a) is able to develop asymptotic distribution of Cox-type tests under a sequence of local alternatives defined in terms of the degree of multicollinearity of the regressors from the two rival models. Under this sequence of local alternatives he shows that the F test based on the comprehensive model is less powerful than the Cox-type tests, unless the number of non-overlapping variables of the alternative over the null hypothesis is unity. An alternative approach to asymptotic power comparisons which does not require specification of local alternatives is advanced by Bahadur (1960) and Bahadur (1967) and holds the alternative hypothesis fixed but allows the size of the test to tend to zero as sample size increases. Asymptotic power comparisons of non-nested tests by the Bahadur approach is considered in Gourieroux (1982) and Pesaran (1984).

6. Measures of Closeness and Vuong's Approach

So far the concepts of nested and non-nested hypotheses have been loosely defined, but for a more integrated approach to non-nested hypothesis testing and model selection a more formal definition is required. This can be done by means of a variety of "closeness" criteria proposed in the literature for measuring the divergence of one distribution function with respect to another. A popular measure employed in Pesaran (1987) for this purpose is the Kullback-Leibler (Kullback (1959)) Information Criterion (KLIC). This criteria has been used extensively in the development of both non-nested hypotheses tests and model selection procedures. Given hypotheses H_f

[22] See also McAleer (1983).

[23] See McAleer and Pesaran (1986) for additional details.

and H_g, defined by (2.4) and (2.5), the KLIC measure of H_g with respect to H_f is written as

$$\begin{aligned} I_{fg}(\boldsymbol{\theta},\boldsymbol{\gamma}) &= E_f\{\ln f_t(\boldsymbol{\theta}) - \ln\ g_t(\boldsymbol{\gamma})\} \\ &\int_{R_f} \ln\left\{\frac{f_t(\boldsymbol{\theta})}{g_t(\boldsymbol{\gamma})}\right\} f_t(\boldsymbol{\theta})dy. \end{aligned} \tag{6.1}$$

It is important to note that $I_{fg}(\boldsymbol{\theta},\boldsymbol{\gamma})$ is not a distance measure. For example, in general $I_{fg}(\boldsymbol{\theta},\boldsymbol{\gamma})$ is not the same as $I_{gf}(\boldsymbol{\gamma},\boldsymbol{\theta})$, and KLIC does not satisfy the triangular inequality, namely $I_{fg} + I_{gh}$ need not exceed I_{fh} as required if KLIC were a distance measure. Nevertheless, KLIC has a number attractive properties: $I_{fg}(\boldsymbol{\theta},\boldsymbol{\gamma}) \geq 0$, with the strict equality holding if and only if $f(.) = g(.)$. Assuming that observations on y_t are independently distributed then the KLIC measure is additive over sample observations.

To provide a formal definition of non-nested or nested hypothesis we define two "closeness" measures: one measuring the closeness of H_g to H_f (viewed from the perspective of H_f), and another the closeness measure of H_f to H_g. These are respectively given by $C_{fg}(\boldsymbol{\theta}_0) = I_{fg}(\boldsymbol{\theta}_0, \boldsymbol{\gamma}_*(\boldsymbol{\theta}_0))$, and $C_{gf}(\boldsymbol{\gamma}_0) = I_{gf}(\boldsymbol{\gamma}_0, \boldsymbol{\theta}_*(\boldsymbol{\gamma}_0))$, where as before $\boldsymbol{\gamma}_*(\boldsymbol{\theta}_0)$ is the pseudo-true value of $\boldsymbol{\gamma}$ under H_f, and $\boldsymbol{\theta}_*(\boldsymbol{\gamma}_0)$ is pseudo-true value of $\boldsymbol{\theta}$ under H_g.

Definition 6.1. *H_f is nested within H_g if and only if $C_{fg}(\boldsymbol{\theta}_0) = 0$, for all values of $\boldsymbol{\theta}_0 \in \Theta$, and $C_{gf}(\boldsymbol{\gamma}_0) \neq 0$ for some $\boldsymbol{\gamma}_0 \in \Gamma$.*

Definition 6.2. *H_f and H_g are globally non-nested if and only if $C_{fg}(\boldsymbol{\theta}_0)$ and $C_{gf}(\boldsymbol{\gamma}_0)$ are both non-zero for all values of $\boldsymbol{\theta}_0 \in \Theta$ and $\boldsymbol{\gamma}_0 \in \Gamma$.*

Definition 6.3. *H_f and H_g are partially non-nested if $C_{fg}(\boldsymbol{\theta}_0)$ and $C_{gf}(\boldsymbol{\gamma}_0)$ are both non-zero for some values of $\boldsymbol{\theta}_0 \in \Theta$ and $\boldsymbol{\gamma}_0 \in \Gamma$.*

Definition 6.4. *H_f and H_g are observationally equivalent if and only if $C_{fg}(\boldsymbol{\theta}_0) = 0$ and $C_{gf}(\boldsymbol{\gamma}_0) = 0$ for all values of $\boldsymbol{\theta}_0 \in \Theta$ and $\boldsymbol{\gamma}_0 \in \Gamma$.*

Using the above definitions it is easily seen, for example, that linear or non-linear rival regression models can at most be partially non-nested, but exponential and log-normal distributions discussed in Section 3 are globally non-nested. For further details see Pesaran (1987).

We may also define a closeness measure of H_g to H_f from the perspective of the true model H_h and in doing so are able to motivate Vuong's approach to hypothesis testing and model selection. (see Vuong (1989)). The primary focus of Vuong's analysis is to test the hypothesis that the models under consideration are "equally" close to the true model. As Vuong (1989) notes *"If the distance between a specified model and the true distribution is defined as the minimum of the KLIC over the*

distributions in the model, then it is natural to define the "best" model among a collection of competing models to be the model that is closest to the true distribution". Thus, in contrast to the standard approach to model selection, a hypothesis testing framework is adopted and *a probabilistic* decision rule used to select a "best" model.

With our set up and notations the closeness of H_f to H_h viewed from the perspective of the true model, H_h is defined by

$$C_{hf}(\boldsymbol{\theta}_{h*}) = E_h\{\ln h_t(\cdot) - \ln f_t(\boldsymbol{\theta}_{h*})\}.$$

Similarly, the closeness of H_g to H_h is defined by

$$C_{hg}(\boldsymbol{\gamma}_{h*}) = E_h\{\ln h_t(\cdot) - \ln g_t(\boldsymbol{\gamma}_{h*})\}.$$

The null hypothesis underlying Vuong's approach is now given by

$$H_V : \; C_{hf}(\boldsymbol{\theta}_{h*}) = C_{hg}(\boldsymbol{\gamma}_{h*}),$$

which can also be written as

$$H_V : \; E_h\{\ln f_t(\boldsymbol{\theta}_{h*})\} = E_h\{\ln g_t(\boldsymbol{\gamma}_{h*})\}.$$

The quantity $E_h\{\ln f_t(\boldsymbol{\theta}_{h*}) - \ln g_t(\boldsymbol{\gamma}_{h*})\}$ is unknown and depends on the unknown true distribution H_h, but can be consistently estimated by the average log-likelihood ratio statistic, $T^{-1}\left\{L_f(\hat{\boldsymbol{\theta}}_T) - L_g(\hat{\boldsymbol{\gamma}}_T)\right\}$. Vuong derives the asymptotic distribution of the average log-likelihood ratio under H_V, and shows that it crucially depends on whether $f_t(\boldsymbol{\theta}_{h*}) = g_t(\boldsymbol{\gamma}_{h*})$, namely whether the distributions in H_f and H_g that are closest to the true model are observationally equivalent or not. In view of this a sequential approach to hypothesis testing is proposed. See Vuong (1989) for further details.

7. Practical Problems

In Section 5 we noted that the motivation for the Cox test statistic was based upon the observation that unless two models, say $f(.)$ and $g(.)$ are non-nested then the expectation

$$T^{-1}E_f\left[L_f(\boldsymbol{\theta}) - L_g(\boldsymbol{\gamma})\right], \tag{7.1}$$

does not evaluate to zero and as a result standard likelihood ratio statistics are not appropriate. Cox (1961,1962) proposed a procedure such that a centred (modified) log-likelihood ratio has a well-defined limiting distribution. In Section 5.1 we demonstrated that in the case of the linear regression we may obtain a closed form consistent estimate of (7.1). However, this is the exception rather than the rule and the use of the Cox test has been restricted to a relatively small number of

applications due to problems in constructing a consistent estimate of the expected log-likelihood ratio statistic. There are two principal problems. First, in order to estimate (7.1) we require a consistent estimate of the pseudo true value, $\boldsymbol{\gamma}(\boldsymbol{\theta}_0)$. Second, in most cases even given such an estimate, the expectation (7.1) will still be intractable. An exception is the application of the Cox test to both binary and multinomial probit and logit models. Independent of the dimension of the choice set, the expected difference between the two log-likelihoods under the null has a relatively simple, closed form expression (see Pesaran and Pesaran (1993)).

Following the work of Pesaran and Pesaran (1993), Pesaran and Pesaran (1995), and Weeks (1996), a simulation-based application of the modified likelihood principle has been used to affect adjustments to the test statistic in order to improve the finite sample size and power properties. A drawback of this approach is that it is still reliant upon a reference distribution which is valid asymptotically. In addition, Orme (1994) attests to the existence of a large number of asymptotically equivalent (AE) variants of the Cox test statistic which represents a formidable menu of choices for the applied econometrician. In the case of the numerator, various test statistics are based upon the use of alternative consistent estimators of the Kullback-Leibler measure of closeness. An additional set of variants of the Cox test statistic depend upon the existence of a number of AE ways of estimating the variance of the test statistic.

An alternative approach based upon the seminal work of Efron (1979), with contributions by Hall (1986), Beran (1988), Hinkely (1988), and Coulibaly and Brorsen (1998), applies bootstrap-based procedures to directly evaluate the empirical distribution function of the log-likelihood ratio statistic. In this context the focus is upon correcting the reference distribution rather than centring the log-likelihood ratio statistic and utilising limiting distribution arguments. This type of adjustment may, in a number of cases, be theoretically justified through Edgeworth expansions and can under certain conditions result in improvements over classical asymptotic inference. The existence of a large menu of broadly equivalent test statistics is also relevant in the context of bootstrap-based inference. Recent surveys by Vinod (1993), Jeong and Maddala (1993), and Li and Maddala (1996), review a large number of variants including the double, recursive and weighted bootstrap. Related, Hall (1988) notes that in many applications the precise nature of the bootstrap design is not stated.

7.1. A Simulation Application of the Modified Likelihood Principle

The essence of the Cox non-nested test is that the mean adjusted ratio of the maximised log-likelihoods of two non-nested models has a well defined *limiting* distribution under the null hypothesis. Using the notation set out in Section 2 above we may write the numerator of the Cox test statistic as

$$S_{fg} = T^{-1} LR_{fg} - C_{fg}(\hat{\theta}_T, \widetilde{\boldsymbol{\gamma}}). \tag{7.2}$$

The last term on the right-hand side of (7.2), $C_{fg}(\hat{\theta}_T, \tilde{\boldsymbol{\gamma}})$, represents a consistent estimator of $C_{fg}(\boldsymbol{\theta}_0, \boldsymbol{\gamma}_*(\boldsymbol{\theta}_0))$, the KLIC measure of closeness of $g(.)$ to $f(.)$. This may be written as $C_{fg}(\hat{\theta}_T, \widetilde{\boldsymbol{\gamma}}) = \hat{E}_f \left[T^{-1}(L_f(\hat{\theta}_T) - L_g(\widetilde{\boldsymbol{\gamma}})) \right]$, and is an estimator of the difference between the expected value of the two maximised log-likelihoods under the distribution given by $f(.)$; $\widetilde{\boldsymbol{\gamma}}$ is any consistent estimator for $\boldsymbol{\gamma}_*(\boldsymbol{\theta}_0)$. Weeks (1996) in testing probit and logit models of discrete choice, distinguished between three variants, $\widetilde{\boldsymbol{\gamma}} = \{\widehat{\boldsymbol{\gamma}}_T, \boldsymbol{\gamma}_R(\widehat{\boldsymbol{\theta}}_T), \overline{\boldsymbol{\gamma}}_T\}$. $\hat{\boldsymbol{\gamma}}_T$ is the MLE of $\boldsymbol{\gamma}$, $\overline{\boldsymbol{\gamma}}$ is due to Kent (1986) and is an estimator derived from maximising the fitted log-likelihood, and $\boldsymbol{\gamma}_{*R}(\widehat{\boldsymbol{\theta}}_T) = \frac{1}{R}\sum_{r=1}^{R} \boldsymbol{\gamma}_*^r(\widehat{\boldsymbol{\theta}}_T)$ is a simulation-based estimator where $\boldsymbol{\gamma}_*^r(\widehat{\boldsymbol{\theta}}_T)$ is the solution to

$$\underset{\gamma}{\text{Arg max}}\{L_g^r(\boldsymbol{\gamma}) = \sum_{t=1}^{T} \ln g(\mathbf{y}_t^r(\hat{\boldsymbol{\theta}}_T)|\mathbf{z}_t, \Omega_{t-1}; \boldsymbol{\gamma})\}, \tag{7.3}$$

where $\mathbf{y}_t^r(\hat{\boldsymbol{\theta}}_T)$ is the *rth* draw of $\mathbf{y}_t$ under H_f using $\widehat{\boldsymbol{\theta}}_T$ and R is the number of simulations. Note that for both $R \to \infty$ and $T \to \infty$ then $\boldsymbol{\gamma}_{*R}(\widehat{\boldsymbol{\theta}}_T) \to \boldsymbol{\gamma}_*(\boldsymbol{\theta}_0)$.

A simulation-based estimator of $C_{fg}(\boldsymbol{\theta}_0, \boldsymbol{\gamma}_*(\boldsymbol{\theta}_0))$ has been suggested by Pesaran and Pesaran (1993) and is given by

$$C_{fg,R}(\hat{\boldsymbol{\theta}}_T, \boldsymbol{\gamma}_{*R}(\hat{\boldsymbol{\theta}}_T)) = \frac{1}{TR}\sum_{r=1}^{R} \left[L_f^r(\hat{\boldsymbol{\theta}}_T) - L_g^r(\boldsymbol{\gamma}_{*R}(\hat{\boldsymbol{\theta}}_T)) \right]. \tag{7.4}$$

However (7.4) represents one approach to centring the log-likelihood ratio statistic, whereby both $\hat{\boldsymbol{\theta}}_T$ and $\boldsymbol{\gamma}_{*R}(\hat{\boldsymbol{\theta}}_T)$ are treated as *fixed* parameters. An alternative method of mean adjustment is given by the following estimator of KLIC

$$C_{fg,R}(\hat{\boldsymbol{\theta}}_T^1, \ldots, \hat{\boldsymbol{\theta}}_T^R, \boldsymbol{\gamma}_*^1(\hat{\boldsymbol{\theta}}_T), \ldots, \boldsymbol{\gamma}_*^R(\hat{\boldsymbol{\theta}}_T)) = \frac{1}{TR}\sum_{r=1}^{R} \left[L_f^r(\hat{\boldsymbol{\theta}}_T^r) - L_g^r(\boldsymbol{\gamma}_*^r(\hat{\boldsymbol{\theta}}_T)) \right], \tag{7.5}$$

where the parameter arguments to both $L_f(.)$ and $L_g(.)$ are allowed to *vary* across each *rth* replication. (See Coulibaly and Brorsen (1998)).

7.2. Resampling the Likelihood Ratio Statistic: Bootstrap Methods

The bootstrap is a data-based simulation method for statistical inference. The bootstrap approach involves approximating the distribution of a function of the observed data by the bootstrap distribution of the quantity. This is done by substituting the empirical distribution function for the unknown distribution and repeating this process many times to obtain a simulated distribution. Its recent development follows from the requirement of a significant amount of computational power. Obviously there is no advantage to utilising bootstrap procedures when the exact sampling distribution of the test statistic is known. However, it has been demonstrated that when the sampling distribution is not known, the substitution of computational intensive

bootstrap resampling can offer an improvement over asymptotic theory. The use of *non-pivotal* bootstrap testing procedures does not require the mean adjustment facilitated by (7.4) and (7.5). However, pivotal (or bootstrap-t) procedures require both mean and variance adjustments in order to guarantee asymptotic pivotalness.

Utilising a parametric bootstrap we present below a simple algorithm for re-sampling the likelihood ratio statistic which we then use to construct the empirical distribution function of the test statistic. For the purpose of exposition the algorithm is presented for the non-pivotal bootstrap.

1. Generate R samples of size T by sampling from the *fitted* null model $f_t(\hat{\boldsymbol{\theta}}_T)$.

2. For each rth simulated sample, the pair $(\hat{\boldsymbol{\theta}}_T^r, \boldsymbol{\gamma}_*^r(\hat{\boldsymbol{\theta}}_T))$ represent the parameter estimates obtained by maximising the log likelihoods

$$L_f^r(\boldsymbol{\theta}) = \sum_{t=1}^{T} \ln f_t(\mathbf{y}_t^r(\hat{\boldsymbol{\theta}}_T)|\mathbf{x}_t, \Omega_{t-1}; \boldsymbol{\theta}), \;\; L_g^r(\boldsymbol{\gamma}) = \sum_{t=1}^{T} \ln g_t(\mathbf{y}_t^r(\hat{\boldsymbol{\theta}}_T)|\mathbf{z}_t, \Omega_{t-1}; \boldsymbol{\gamma}), \tag{7.6}$$

   where $\mathbf{y}_t^r(\hat{\boldsymbol{\theta}}_T)$ denotes the rth bootstrap-sample conditional upon $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_T$. We then compute the simulated log likelihood ratio statistic

$$T_f^r = L_f(\hat{\boldsymbol{\theta}}_T^r) - L_g(\boldsymbol{\gamma}_*^r(\hat{\boldsymbol{\theta}}_T)).$$

3. By constructing the empirical cdf of $\{T_f^r : 1 \leq r \leq R\}$, we can compare the *observed* test statistic, $T_f = L_f(\hat{\boldsymbol{\theta}}_T) - L_g(\boldsymbol{\gamma}_*(\hat{\boldsymbol{\theta}}_T))$, with critical values obtained from the R independent (conditional) realisations of T_f^r. The p-value based upon the bootstrap procedure is given by[24]

$$P_R = \frac{\sum_{r=1}^{R} \mathbf{1}(T_f^r \geq T_f)}{R}, \tag{7.7}$$

   where $\mathbf{1}(.)$ is the indicator function.

The bootstrap procedure outlined above simply resamples the likelihood ratio statistic *without* pivoting. There are a number of alternative test statistics which by using pivotal methods are conjectured to represent an improvement over classical first order methods (see for example, Beran (1988) and Hall (1988)). An evaluation of both the size and power properties of a number of simulation and bootstrap-based tests applied to linear versus loglinear regression models and a number of variants of threshold autoregressive models is provided in Kapetanios and Weeks (1999).

[24] If T is discrete then repeat values of T can occur requiring that we make an adjustment to (7.7).

References

Akaike, H. (1973): "Information Theory and an Extension of the Maximum Likelihood Principle," in *Proceedings of the 2nd International Symposium on Information Theory*, ed. by N. Petrov, and F. Csadki, pp. 267–281. Akademiai Kiado, Budapest.

Altissimo, F., and G. L. Violante (1998): "The Nonlinear Dynamics of Output and Unemployment in the US," Department of Economics, University College, London.

Amemiya, T. (1985): *Advanced Econometrics*. Harvard University Press, Cambridge, M.A.

Aneuryn-Evans, G., and A. S. Deaton (1980): "Testing Linear versus Logarithmic Regression Models," *Review of Economic Studies*, 47, 275–291.

Atkinson, A. (1970): "A Method for Discriminating Between Models (with Discussion)," *Journal of the Royal Statistical Society, B*, B32, 323–353.

Bahadur, R. R. (1960): "Stochastic Comparison of Tests," *Ann. Math. Statist.*, 31, 276–95.

——— (1967): "Rates of Convergence of Estimates and Test Statistics," *Ann. Math. Statist.*, 38, 303–24.

Barro, R. (1977): "Unanticipated Money Growth and Unemployment in the United States," *American Economic Review*, 67, 101–15.

Beran, R. (1988): "Prepivoting Test Statistics: A Bootstrap View of Asymptotic Refinements," *Journal of the American Statistical Association*, 83(403).

Bernanke, B., H. Bohn, and P. Reiss (1988): "Alternative Non-Nested Specification Tests of Time-Series Investment Models," *Journal of Econometrics*, 37, 293–326.

Chao, J. C., and N. R. Swanson (1997): "Tests of Non-Nested Hypotheses in Nonstationary Regressions with an Application to Modeling Industrial Production," Working Paper, Department of Economics, Pennsylvania State University.

Coulibaly, N., and B. Brorsen (1998): "A Monte Carlo Sampling Approach to Testing Nonnested Hypotheses: Monte Carlo Results," *Econometric Reviews*, pp. 195–209.

Cox, D. (1961): "Tests of Separate Families of Hypothesis," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*.

——— (1962): "Further Results on Tests of Separate Families of Hypotheses," *Journal of Royal Statistical Society*, B24, 406–424.

Dastoor, N. (1981): "A Note on the Interpretation of the Cox Procedure for Non-Nested Hypotheses," *Economics Letters*, 8, 113–119.

Dastoor, N. K. (1983): "Some Aspects of Testing Non-Nested Hypotheses," *Journal of Econometrics*, 21, 213–228.

Davidson, R., and J. MacKinnon (1981): "Several Tests for Model Specification in the Presence of Alternative Hypotheses," *Econometrica*, 49, 781–793.

Davidson, R., and J. G. MacKinnon (1985): "Testing Linear and Log-Linear Regressions Against Box-Cox Alternatives," *Canadian Journal of Economics*, 18, 499–517.

Davidson, R., and J. G. McKinnon (1982): "Some Non-Nested Hypothesis Tests and the Relations Among Them," *Review of Economic Studies*, 49, 551–565.

Davies, R. B. (1977): "Hypothesis Testing when a Nuisance Parameter is Present Only under the Alternative," *Biometrika*, 64(2), 247–254.

Deaton, A. S. (1982): "Model Selection Procedures, or, Does the Consumption Function Exist?," in *Evaluating the Reliability of Macroeconomic Models*, ed. by G. C. Show, and P. Corsi, pp. 43–65. Wiley: New York.

Dixit, A. V., and R. S. Pindyck (1994): *Investment Under Uncertainty*. Princeton University Press, Chichester, UK.

Duncan, A., and M. Weeks (1998): "Non-Nested Models of Labour Supply with Discrete Choices," Working Paper, Department of Economics, University of York.

Efron, B. (1979): "Bootstrap Methods: Another Look at the Jacknife," *Annals of Statistics*, 7, 1–26.

Ericsson, N. (1983): "Asymptotic Properties of Instrumental Variables Statistics for Testing Non-Nested Hypotheses," *Review of Economic Studies*, 50, 287–304.

Fisher, G., and M. McAleer (1979): "On the Interpretation of the Cox Test in Econometrics," *Economic Letters*, 4, 145–50.

Fisher, G. R., and M. McAleer (1981): "Alternative Procedures and Associated Tests of Significance for Non-Nested Hypotheses," *Journal of Econometrics*, 16, 103–119.

FRIEDMAN, M., AND D. MEISELMAN (1963): "The Relative Stability of Monetary Velocity and the Investment Multiplier in the United States 1897-1958," in *Stabilization Policies*. Commission on Money and Credit Research Study, Englewood Cliffs, New Jersey.

GHYSELS, E., AND A. HALL (1990): "Testing Non-Nested Euler Conditions with Quadrature Based Method of Approximation," *Journal of Econometrics*, 46, 273–308.

GODFREY, L. G. (1983): "Testing Non-Nested Models After Estimation by Instrumental Variables or Least Squares," *Econometrica*, 51(2), 355–365.

GODFREY, L. G., AND M. H. PESARAN (1983): "Tests of Non-Nested Regression Models: Small Sample Adjustments and Monte Carlo Evidence," *Journal of Econometrics*, 21, 133–154.

GODFREY, L. G., AND M. WICKENS (1981): "Testing Linear and Log-Linear Regressions for Functional Form," *Review of Economic Studies*, 48, 487–496.

GOURIEROUX, C. (1982): "Asymptotic Comparison of Tests for Non-Nested Hypotheses by Bahadur's A.R.E.," Discussion Paper 8215, CEPREMAP, Paris.

GOURIEROUX, C., AND A. MONFORT (1994): "Testing Non-Nested Hypotheses," in *Handbook of Econometrics, Volume IV*, ed. by R. F. Engle, and D. L. McFadden. Elsevier, Oxford.

GOURIEROUX, C., AND A. MONFORT (1995): "Testing, Encompassing, and Simulating Dynamic Econometric Models," *Econometric Theory*, 11, 195–228.

GOURIEROUX, C., A. MONFORT, AND A. TROGNON (1983): "Testing Nested or Non-Nested Hypotheses," *Journal of Econometrics*, 21, 83–115.

GRANGER, C., M. L. KING, AND H. WHITE (1995): "Comments on Testing Economic Theories and the Use of Model Selection Criteria," *Journal of Econometrics*, 67, 173–187.

GRANGER, C. W. J., AND M. H. PESARAN (1999): "A Decision Theoretic Approach to Forecast Evaluation," DAE, Working Paper No. 9618.

GRASA, A. A. (1989): *Econometric Model Selection: A New Approach*. Kluwer Academic Publishers, Spain.

HALL, P. (1986): "On the Number of Bootstrap Simulations Required to Construct a Confidence Interval," *The Annals of Statistics*, 14(4).

——— (1988): "Theoretical Comparison of Bootstrap Confidence Intervals," *Annals of Statistics*, 16, 927–953.

Hendry, D. F. (1993): *Econometrics: Alchemy or Science?* Blackwell Publishers, Oxford.

Hinkely, D. (1988): "Bootstrap Methods," *Journal of the Royal Statistical Society (Series B)*, 50, 321–337.

Jeong, J., and G. Maddala (1993): "A Perspective on Application of Bootstrap Methods in Econometrics," *Handbook of Statistics*, 11, 573–605.

Jorgenson, D. W., and C. D. Siebert (1968): "A Comparison of Alternative Theories of Corporate Investment Behavior," *American Economic Review*, 58, 681–712.

Kapetanios, G., and M. Weeks (1999): "Non-Nested Models and the Likelihood Ratio Statistic: A Comparison of Simulation and Bootstrap-Based Tests," Department of Applied Economics Working Paper.

Kent, J. (1986): "The Underlying Structure of Nonnested Hypothesis Tests," *Biometrika*, 7, 333–43.

King, M. L. (1983): "Testing for Autoregressive Against Moving Average Errors in the Linear Regression Model," *Journal of Econometrics*, 21, 35–51.

Kullback, S. (1959): *Information Theory and Statistics*. Wiley, New York.

Lavergne, P. (1998): "Selection of Regressors in Econometrics: Parametric and Nonparametric Methods," *Econometric Reviews*, 17(3), 227–273.

Leamer, E. E. (1983): "Model Choice and Specification Analysis," in *Handbook of Econometrics*, ed. by Z. Griliches, and M. D. Intriligator, vol. I. North-Holland Publishing Company, University of California, Los Angeles.

Li, H., and G. Maddala (1996): "Bootstrapping Time Series Models," *Econometric Reviews*, 15(2), 115–158.

Linhart, H., and W. Zucchini (1986): *Model Selection*. Wiley and Sons., New-York.

Maddala, G. S. E. (1981): *Model Selection*. Special Issue of Journal of Econometrics, 16(1).

McAleer, M. (1983): "Exact Tests of a Model Against Non-Nested Alternative," *Biometrika*, 70, 285–288.

McAleer, M. (1995): "The Significance of Testing Empirical Non-Nested Models," *Journal of Econometrics*, 67, 149–171.

McAleer, M., and S. Ling (1998): "A Non-Nested Test for the GARCH and E-GARCH Models," Working Paper, Department of Economics, University of Western Australia.

McAleer, M., and M. H. Pesaran (1986): "Statistical Inference in Non-Nested Econometric Models," *Applied Mathematics and Computation*, 20, 271–311.

McAleer, M. G., G. Fisher, and P. Volker (1982): "Separate Misspecified Regressions and U.S. Long Run Demand for Money Function," *Review of Economics and Statistics*, 64, 572–583.

McAleer, M. J., M. H. Pesaran, and A. K. Bera (1990): "Alternative Approaches to Testing Non-Nested Models with Autocorrelated Disturbances: An Application to Models of U.S. Unemployment," *Communications in Statistics*, series A(19), 3619–44.

Mizon, G. E., and J. F. Richard (1986): "The Encompassing Principle and its Application to Non-Nested Hypotheses," *Econometrica*, 54, 657–678.

Orme, C. (1994): "Non-Nested Tests for Discrete Choice Models," Working Paper, Dept. of Economics, University of York.

Pereira, B. d. B. (1984): "On the Choice of a Weibull Model," *Journal of the Inter American Statistical Institute*, 26, 157–163.

Pesaran, M. H. (1974): "On the General Problem of Model Selection," *Review of Economic Studies*, 41, 153–171.

——— (1981): "Pitfalls of Testing Non-Nested Hypotheses by the Lagrange Multiplier Method," *Journal of Econometrics*, 17, 323–331.

——— (1982a): "Comparison of Local Power of Altenative Tests of Non-Nested Regression Models," *Econometrica*.

——— (1982b): "A Critique of the Proposed Tests of the Natural Rate-Rational Expectations Hypothesis," *The Economic Journal*, 92, 529–554.

——— (1984): "Asymptotic Power Comparisons of Tests of Separate Parametric Families by Bahadur's Approach," *Biometrika*, 71(2), 245–52.

——— (1987): "Global and Partial Nonnested Hypotheses and Asymptotic Local Power.," *Econometric Theory*, 3, 69–97.

Pesaran, M. H., and S. Deaton (1978): "Testing Non-Nested Non-Linear Regression Models," *Econometrica*, 46, 677–694.

Pesaran, M. H., and B. Pesaran (1993): "A Simulation Approach to the Problem of Computing Cox's Statistic for Testing Non-Nested Models," *Journal of Econometrics*, 57, 377–92.

——— (1995): "A Non-Nested Test of Level Differences Versus Log-Differenced Stationary Models," *Econometric Reviews*, 14(2), 213–27.

Pesaran, M. H., and S. Potter (1997): "A Floor and Ceiling Model of US Output," *Journal of Economic Dynamics and Control*, 21(4-5), 661–696.

Sawyer, K. R. (1983): "Testing Separate Families of Hypotheses: An Information Criterion," *Journal of the Royal Statistical Society B*, 45, 89–99.

——— (1984): "Multiple Hypothesis Testing," *Royal Statistical Society B*, 46(3), 419–424.

Schwarz, G. (1978): "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461–464.

Smith, M. A., and G. S. Maddala (1983): "Multiple Model Testing for Non-Nested Heteroskedastic Censored Regression Models," *Journal of Econometrics*, 21, 71–81.

Smith, M. D., and D. J. Smyth (1991): "Multiple and Pairwise Non-Nested Tests of the Influece of Taxes on Money Demand," *Journal of Applied Econometrics*, 6, 17–30.

Smith, R. J. (1992): "Non-Nested Tests for Competing Models Estimated by Generalized Method of Moments," *Econometrica*, 60, 973–980.

——— (1993): "Consistent Tests for the Encompassing Hypothesis," Document de Travail No. 9403, INSEE, Paris.

Vinod, H. (1993): "Bootstrap Methods: Applications in Econometrics," *Handbook of Statistics*, 11, 629–661.

Vuong, Q. H. (1989): "Likelihood Ratio Tests for Model Selection and Non-Nested Hypothesis," *Econometrica*, 57(2), 307–333.

Walker, A. M. (1967): "Some Tests of Separate Families of Hypotheses in Time Series Analysis," *Biometrika*, 54, 39–68.

Weeks, M. (1996): "Testing the Binomial and Multinomial Choice Models Using Cox's Non-Nested Test," *Journal of the American Statistical Association (Papers and Proceedings)*.

White, H. (1982): "Regularity Conditions for Cox's Test of Nonnested Hypothesis," *Journal of Econometrics*, 19, 301–18.

Zellner, A. (1971): *An Introduction to Bayesian Inference in Econometrics*. John Wiley and Sons, New York.