SELECTING THE BEST LINEAR REGRESSION MODEL:
A CLASSICAL APPROACH

Donald Lien and Quang H. Vuong
California Institute of Technology

SOCIAL SCIENCE WORKING PAPER 606
March 1986
In this paper, we apply the model selection approach based on Likelihood Ratio (LR) tests developed in Vuong (1985) to the problem of choosing between two normal linear regression models which are not nested in each other. First we compare our model selection procedure to other model selection criteria. Then we explicitly derive the procedure when the competing linear models are non-nested and neither one is correctly specified. Some simplifications are seen to arise when both models are contained in a larger correctly specified linear regression model, or when at least one competing linear model is correctly specified. A comparison of our model selection tests and previous non-nested hypothesis tests concludes the paper.
SELECTING THE BEST LINEAR REGRESSION MODEL:
A CLASSICAL APPROACH

Donald Lien and Quang H. Vuong
California Institute of Technology

1. INTRODUCTION

In this paper, we apply the model selection approach developed in Vuong (1985) to the classical problem of choosing between two linear regression models. That this problem is important to applied econometricians results from the fact that one does not in general have a unique econometric model either because one wants to compare many theories or because a theory does not provide a unique functional form.

The problem of selecting the "best" subset of variables in a linear regression context has long been of special interest to theoretical and applied statisticians. The numerous papers that were generated by this classical problem have recently been surveyed by Gaver and Geisel (1974), Hocking (1976), and Lindley (1968) among others. Various solutions dealing with different aspects of this problem were proposed.

A first general approach is to set the problem in a decision framework. The natural solution is then the Bayesian solution which relies on Jeffrey's posterior probability criterion (see in particular Zellner (1971)). This solution will not be discussed in this paper. Other solutions in a decision theoretic framework but which are not always justified in a Bayesian setting are based on the construction of model selection criteria. Most of these widely used criteria will be discussed and compared to our solution later in this paper.

A second general approach is to adopt the classical hypothesis testing framework. In this context, two solutions are in general accepted. The first solution derives from the work of Cox (1960, 1961) on testing non-nested hypotheses. Starting with Pesaran (1974), this solution has recently attracted a lot of attention from theoretical econometricians. The second solution consists in nesting the competing models in a larger model and to test that the additional parameters are equal to some particular values (Atkinson (1970)).

Within the classical hypothesis framework, there is however a third solution which has not been widely recognized and which dates back to Hotelling (1940). It consists in discriminating between the competing models by testing the hypothesis that the models are "equivalent" under some appropriate definition. Recent works along this line are White and Olson (1979) where the mean squared error of prediction is used, and Vuong (1985) where the Kullback-Leibler (1951) criterion is used. The advantage of this discriminating approach is that, unlike other classical solutions, the competing models are treated symmetrically.

The purpose of this paper is to develop this discriminating solution for the case where the competing models are normal linear regressions. Since neither model may be correctly specified, by necessity, this paper is mainly concerned with asymptotic results. The paper is organized as follows. In Section 2, we discuss two theoretical model selection criteria, i.e., mean squared error (MSE)
of prediction and Kullback-Leibler information criterion (KLIC). They represent two different distance measures between two probability distributions. Upon applying the two criteria to normal linear regression models, we show that they lead to the same comparison. In Section 3, we show that most of the model selection criteria in the literature are either consistent estimates of MSE or consistent estimates of KLIC. Based on this remark, a short survey is provided.

We then turn to the model selection approach based on Likelihood Ratio (LR) tests that is developed in Vuong (1985). Specifically, we characterize this procedure when the competing linear regression models are non-nested and neither one is correctly specified. The complicated results are presented in Section 4. Some simplifications are seen to arise when both models are contained in a larger linear regression model which is correctly specified or when at least one model is correctly specified. These results are discussed in Section 5 and Section 6, respectively. A comparison of our model selection tests and previous non-nested hypothesis tests concludes the paper. All the proofs are collected in the Appendix.

2. TWO THEORETICAL MODEL SELECTION CRITERIA

In this section we consider two important theoretical model selection criteria and we apply them to the normal linear regression model. We shall argue in the next section that most of the current model selection criteria can be thought of as estimates of either one of these two theoretical criteria. Unlike previous studies on model selection that assume fixed explanatory variables or fixed in repeated sample, we shall assume that our explanatory variables are random, an assumption that is justified with economic data. Let \((y_t, x_t')\) be the t-th observation on the \((1 + f)\)-dimensional random vector \((y, x')\) defined on an Euclidean measurable space. For simplicity, we adopt Vuong (1983) framework and assume:

**Assumption A1:** The random vectors \((y_t, x_t')\), \(t = 1, 2, \ldots\) are independent and identically distributed with common true cumulative distribution function \(H_0, 1\).

In econometric modelling, we are interested in the true conditional distribution of \(y\) given \(x\). Let \(H_0 (\cdot | \cdot , \cdot)\) denote such a conditional distribution. To estimate \(H_0 (\cdot | \cdot , \cdot)\), we specify parametric conditional models for \(y\) given \(x\), i.e., parametric families of conditional distributions for \(y\) given \(x\), \(F_{\theta} = \{F_{\theta} (\cdot | \cdot , \cdot) \in \Theta \} \subseteq \{y \mid x\} \text{ conditional distribution} \). In this paper, we shall consider linear regression models with normal errors. Each linear regression model will be associated with a subset of the “exogenous” variables \(x\). Specifically, let \(x_s\) be a \(f_s\)-subset of \(x\). Then, the normal linear regression model for \(y\) given \(x\) with explanatory variables \(x_s\) is formally defined as:

\[
M_s = \{N(\lambda_0 + x_s' \beta_s, \sigma_s^2) ; \theta_s = (\lambda_0, \beta_s, \sigma_s^2) \in \mathbb{R}^{f_s+1} \times \mathbb{R}_+ \} \quad (2.1)
\]
where \( N(\mu, \sigma^2) \) denotes the univariate normal distribution with mean \( \mu \) and variance \( \sigma^2 \). Thus the \( \Theta \)-conditional distribution for \( y \) given \( x \) in \( M_\Theta \) specifies that \( E_\Theta(y|x) = \lambda_0 + x_0 \lambda_\Theta \) and \( \text{Var}_\Theta(y|x) = \sigma^2_\Theta \).

To evaluate the adequacy of a specified conditional model for \( y \) given \( x \), it is first necessary to define a measure of distance between the true conditional distribution \( H^0 \) and a given conditional distribution \( F_\Theta (x) \). Two measures of distance are generally accepted.

The first measure is based on the mean squared error (MSE) of prediction. Let \( E_\Theta(y|x) \) be the conditional expectation of \( y \) given \( x \) for the conditional distribution \( F_\Theta (x) \), i.e.,

\[
E_\Theta(y|x) = \int y \, dF_\Theta (x). \tag{2.2}
\]

which is assumed to exist. Then the distance measure based on the mean squared error of prediction is defined as:

\[
\text{MSE}(H^0, F_\Theta (x)) = \text{E}^0 \left[ (y - E_\Theta(y|x))^2 \right]_{y|x} = \text{E}^0 \left[ \text{E}^0 \left[ (y - E_\Theta(y|x))^2 \right]_{x|y} \right]_x \tag{2.3}
\]

where \( \text{E}^0(\cdot) \) indicates that the expectation is evaluated with respect to the true distribution \( H^0 \) of \( (y,x) \). It can easily be shown that an equivalent form for (2.3) is:

\[
\text{MSE}(H^0, F_\Theta (x)) = \text{E}^0 \left[ \text{Var}^0(y|x) \right]_x + \text{E}^0 \left[ (E^0(y|x) - E_\Theta(y|x))^2 \right]_x \tag{2.4}
\]

where the first term is independent of \( \Theta \), and \( \text{E}^0(y|x) \) and \( \text{Var}^0(y|x) \) denote the true conditional expectation and variance of \( y \) given \( x \).

Then the distance between the true conditional distribution \( H^0 \) and a specified conditional model \( F_\Theta (x) \) is defined by:

\[
\text{MSE}(H^0, F_\Theta (x)) = \inf_{\Theta \in \Theta} \text{MSE}(H^0, F_\Theta (x)). \tag{2.5}
\]

Given appropriate regularity conditions (see, e.g., White (1981)), there will exist a \( \Theta^* \) in \( \Theta \) that minimizes \( \text{MSE}(H^0, F_\Theta (x)) \). In such a case, the adequacy of the model \( F_\Theta (x) \) is evaluated by:

\[
\text{MSE}(H^0, F_\Theta (x)) = \text{E}^0 \left[ (y - E^0_\Theta(y|x))^2 \right]_{y|x} = \text{E}^0 \left[ \text{E}^0 \left[ (y - E^0_\Theta(y|x))^2 \right]_{x|y} \right]_x \tag{2.6}
\]

where

\[
\Theta^* = \text{argmin}_{\Theta \in \Theta} \text{E}^0 \left[ (y - E^0_\Theta(y|x))^2 \right]_{y|x} \tag{2.7}
\]

The second measure of adequacy of a conditional model is based on the Kullback-Leibler (1951) Information Criterion (KLIC). Specifically, the distance between the true conditional distribution \( H^0 \) and a given conditional distribution \( F_\Theta (x) \) is defined as:

\[
\text{KLIC}(H^0, F_\Theta (x)) = \text{E}^0 \left[ \log \frac{h^0(y|x)}{f(y|x;\Theta)} \right]_{y|x} \tag{2.8}
\]
where $h^0(\cdot|\cdot)$ and $f(\cdot|\cdot;\theta)$ denote the conditional densities of $H_0$ and $F(\cdot;\theta)$ with respect to a common measure $\nu_y$. In our case, $\nu_y$ will be the Lebesgue measure since $y$ takes its value in $\mathbb{R}$. The conditional density $f(\cdot|\cdot;\theta)$ clearly exists since we are considering normal linear regression models. On the other hand, to ensure the existence of the true density $h^0(\cdot|\cdot)$ we make the following assumption, which will be useful later on. Let $H_0$ be the true marginal distribution of $X$.

**Assumption A2:** For $H_0$-almost all $x$, $H_0(\cdot|x)$ admits a strictly positive density $h^0(\cdot|x)$ with respect to the Lebesgue measure $\nu_x$.

As for the previous distance measure, the distance between the true conditional distribution $H_0$ and a specified conditional model $F_y$ is defined by:

$$KLIC(H_0, F_y) = \inf_{\theta \in \Theta} KLIC(H_0, F_y; \theta)$$

$$= \mathbb{E}^0[\log h^0(x|\cdot)] - \mathbb{E}^0[\log f(y|x;l^\circ)], \quad (2.9)$$

Given appropriate regularity conditions (see, e.g., White (1982a)), there exists a unique $\theta^*$ in $\Theta$, called the pseudo-true parameters (see, e.g., Sawa (1978)), that minimizes $KLIC(H_0, F_y; \theta)$. In this case, the adequacy of a model $F_y$ is evaluated by:

$$KLIC(H_0, F_y) = \mathbb{E}^0[\log h^0(x|\cdot)] - \mathbb{E}^0[\log f(y|x;\theta^*)], \quad (2.10)$$

where

$$\theta^* = \arg\max_{\theta \in \Theta} \mathbb{E}^0[\log f(y|x;\theta)]. \quad (2.11)$$

Each of the above two measures of distance can naturally be used to construct a theoretical model selection criterion. Let $F_\theta$ and $G_\gamma = \{G_y(y); y \in \Gamma\}$ be two competing conditional models for $y$ given $x$. Then, using the mean squared error distance (2.6), we say that:

$$F_\theta \text{ is MSE-better than } G_\gamma \iff \Delta MSE(F_\theta, G_\gamma) > 0,$$

$$F_\theta \text{ is MSE-equivalent to } G_\gamma \iff \Delta MSE(F_\theta, G_\gamma) = 0,$$

$$F_\theta \text{ is MSE-worse than } G_\gamma \iff \Delta MSE(F_\theta, G_\gamma) < 0,$$

where

$$\Delta MSE(F_\theta, G_\gamma) = \mathbb{E}^0[(y - E_x(y|x))^2] - \mathbb{E}^0[(y - E_\gamma(y|x))^2]. \quad (2.12)$$

where $\theta^*$ is defined by (2.7) and $\gamma^*$ by a similar equation for the model $G_\gamma$. Using (2.4), an equivalent expression is:

$$\Delta MSE(F_\theta, G_\gamma) = \mathbb{E}^x[(y-x(y|x))^2] - \mathbb{E}^x[(y-x_\gamma(y|x))^2]. \quad (2.13)$$

Equation (2.13) shows that the definitions of MSE-better, MSE-equivalent, and MSE-worse are in fact identical to those proposed by White and Olson (1979).
Alternatively, using the KLIC distance (2.10), we say that

- $F_\theta$ is KLIC-better than $G_\gamma$ iff $\Delta \text{KLIC}(F_\theta, G_\gamma) > 0$,

- $F_\theta$ is KLIC-equivalent to $G_\gamma$ iff $\Delta \text{KLIC}(F_\theta, G_\gamma) = 0$,

- $F_\theta$ is KLIC-worse than $G_\gamma$ iff $\Delta \text{KLIC}(F_\theta, G_\gamma) < 0$,

where

$$\text{KLIC}(F_\theta, G_\gamma) = \mathbb{E}_0[\log f(y|x; \theta^*)] - \mathbb{E}_0[\log g(y|x, \gamma^*)]$$

and $\theta^*$ and $\gamma^*$ are the pseudo-true parameters for the conditional models $F_\theta$ and $G_\gamma$. The latter definitions were those adopted in Vuong (1985).

The essential difference between the above two sets of definitions follows from the fact that the model selection criterion based on the MSE prediction takes only into account the discrepancy between the true conditional expectation $E^0(y|x)$ and the "best" conditional mean $E_0(y|x)$, while the model selection criterion based on the KLIC takes into account the discrepancy between the whole true conditional density $h^0(\cdot|x)$ and the "best" conditional density $f(\cdot|x; \theta^*)$. Thus a model which is better according to the MSE criterion is not necessarily better according to the KLIC. An important exception, however, is when one conditional model is correctly specified, i.e., when one conditional model contains the true conditional distribution $h^0(\cdot|x)$. Indeed from (2.13), (2.14), and Jensen's inequality, it follows that a correctly specified model is always at least as good as any other models according to either model selection criterion. This latter property is highly desirable and justifies the use of the above model selection criteria in the search of a correctly specified model.

When the competing models are linear regression models with normal errors, the definitions based on the MSE of prediction and on the KLIC are, however, identical as we shall see below. Let $\text{Var}^0(y,x)$ denote the true covariance matrix of $y$ and $x$, which we partition as follows:

$$\text{Var}^0(y,x) = \begin{bmatrix} \sigma_{yy}^0 & \sum_{yx}^0 \\ \sum_{yx}^0 & \sum_{xx}^0 \end{bmatrix}$$

The next assumption rules out perfect multicollinearity among all the exogenous variables $x$.

**Assumption A3:** $\text{Var}^0(y,x)$ is finite, and $\sum_{xx}^0$ is non-singular.

Assumption A3 implies that the true means $\mu_y^0$ and $\mu_x^0$ of $y$ and $x$ exist. The next lemma relates the values of $\theta^*$ defined in (2.7) to the pseudo-true values $\theta^*$ defined in (2.11) when the model is a normal linear regression model with explanatory variables $x_s \subset x$. This follows by noticing that for such a model we have:

$$\mathbb{E}_0[\log d(y|x_s; \theta_s)] = -\frac{1}{2\sigma_s^2} \mathbb{E}_0[\|x_s - \bar{x}_s\|_2^2] = -\frac{1}{2}\log \sigma_s^2 - \frac{1}{2}\log 2\pi. \quad (2.16)$$
Lemma 2.1: Let $M_S$ be a normal linear regression model for $y$ given $x$ with explanatory variables $X_S$. Then, given $A_2 - A_3$,

$$
\lambda_S^* = \lambda_S^+ = \mu_0^+ - \sum_{X_S} (\sum_{X_S})^{-1}\mu_0^+ \quad (2.17)
$$

and

$$
\lambda_S^* = \lambda_S^+ = (\sum_{X_S} X_S X_S)^{-1}\mu_0^+ \quad (2.18)
$$

where $\mu_0^+$, $\sum_{X_S}$, and $\sum_{X_S}$ are the true means and covariances corresponding to the explanatory variables $X_S$.

The next corollary gives a simple interpretation of the pseudo-true values $\theta^*$ under additional assumptions on $H_0$. It is known and stated here for further reference.

Corollary 2.2: In addition to the assumptions of Lemma 2.1, suppose that the true conditional mean $E_0^0(y|x)$ is linear in $x$ and the true conditional variance $\text{Var}_0^0(y|x)$ is independent of $x$, then

$$
E_0^0(y|x) = \lambda_S^0 + X_S^0 \beta_S^0 \quad (2.20)
$$

and

$$
\text{Var}_0^0(y|x) = \sigma_S^2 \quad (2.21)
$$

We are now in a position to establish the equivalence between the MSE criterion and the KLIC for linear regression models. Let $X_f$ and $X_g$ be two subsets (not necessarily disjoint) of $X$. We consider discriminating between two normal linear regression models:

$$
M_f = \{N(\lambda_f^0 + X_f \beta_f, \sigma_f^2) ; \theta_f = (\lambda_f^0, \beta_f, \sigma_f^2) \in \mathbb{R}^{f+1} \times \mathbb{R}_+ \}, \quad (2.22)
$$

$$
M_g = \{N(\lambda_g^0 + X_g \beta_g, \sigma_g^2) ; \theta_g = (\lambda_g^0, \beta_g, \sigma_g^2) \in \mathbb{R}^{g+1} \times \mathbb{R}_+ \}. \quad (2.23)
$$

From Lemma 2.1, we have:

$$
\sigma_f^2 = \sigma_{yy} - \sum_{X_f} (\sum_{X_f})^{-1}\mu_0^+ \quad (2.24)
$$

$$
\sigma_g^2 = \sigma_{yy} - \sum_{X_g} (\sum_{X_g})^{-1}\mu_0^+ \quad (2.25)
$$

Proposition 2.3: Given $A_2 - A_3$,

(i) $\Delta \text{MSE}(M_f, M_g) = \sigma_g^2 - \sigma_f^2 \quad (2.26)$

(ii) $\Delta \text{KLIC}(M_f, M_g) = \frac{1}{2} \log(\sigma_g^2/\sigma_f^2). \quad (2.27)$

Proposition 2.3 shows that, when comparing normal linear regression models, the definitions of better than, equivalent to, and worse than are identical for the MSE criterion and the KLIC.
3. A SURVEY OF SOME MODEL SELECTION PROCEDURES

The quantities $\text{MSE}(H^0(y|x), F_{\hat{\theta}^*})$ and $E^0[\log f(y|x; \hat{\theta}^*)]$, which define the previous two theoretical model selection criteria, are unfortunately unknown. These quantities, which can be viewed as theoretical losses, can nonetheless be consistently estimated. In this section, we shall show that most of the current model selection criteria can be thought of as estimates of either theoretical loss. Our treatment differs from the usual one given in standard textbooks (see, e.g., Chow (1983), Judge et al. (1985)) which introduce these model selection criteria as estimates of the risk

$$R_{n}^{\theta}(\theta^*) = \frac{1}{2} \log 2\pi - \frac{1}{2} \sigma^{2}_{s} - \frac{1}{2} \sigma^{2}_{f} (Y - X_{s}\hat{\theta})'(Y - X_{s}\hat{\theta})$$

where we use the convention that a capital letter denotes a matrix of the $n$ observations on the corresponding random variables. It is well-known that the maximum-likelihood (ML) estimator of $\theta^*$ is given by:

$$\hat{\theta}_{s} = (X_{s}'X_{s})^{-1}X_{s}'y$$

$$\sigma^{2}_{s} = \frac{1}{n-1} \sum_{t=1}^{n} e_{st}^{2}$$

where

$$e_{st} = y_{t} - x_{st}' \hat{\theta}_{s}.$$  

Hence, for $s = f, g$,

$$L_{n}^{\theta}(\theta_{s}) = -(n/2) \log \sigma^{2}_{s} - (n/2) \log 2\pi - n/2.$$  

In addition, from the theory of quasi-ML estimation (see, e.g., White (1982a)), it is known that under A1

$$\hat{\theta}_{s} \xrightarrow{a.s.} \theta_{s}^* \quad \hat{\sigma}^{2}_{s} \xrightarrow{a.s.} \sigma^{2}_{s}, \quad s = f, g.$$  

where $X_{s}$ is a subset of $X$, and $s = f, g$. That is, we exclude the constant term.
whether or not the model \( M_g \) is correctly specified, i.e., whether or not \( H_0: \forall s = f, g \).

We now begin with model selection criteria based on the MSE. From (2.26) and (3.7), it follows that the statistic

\[
\Delta \sigma^2 = \sigma_g^2 - \sigma_f^2
\]

is a consistent estimator of the MSE criterion \( \Delta \text{MSE}(M_f, M_g) \) whether or not the competing models \( M_f \) and \( M_g \) are correctly specified. Such a statistic was proposed by White and Olson (1979) to discriminate between \( M_f \) and \( M_g \) by testing the null hypothesis that the models are (MSE) equivalent.\(^6\)\(^7\)

There exist, however, many other consistent estimators of \( \Delta \text{MSE}(M_f, M_g) \), of which some may have better small sample properties than the natural statistic \( \Delta \sigma^2 \). For instance consider the statistics

\[
\Delta C_p = [\hat{\sigma}_g^2 + \frac{2f_g \hat{\sigma}_g^2}{n - l_f \hat{\sigma}_f^2}] - [\hat{\sigma}_f^2 + \frac{2f_f \hat{\sigma}_f^2}{n - l_f \hat{\sigma}_f^2}]
\]

\[
\Delta P = \frac{n + l_f \hat{\sigma}_f^2}{n - l_f \hat{\sigma}_f^2} - \frac{n + l_f \hat{\sigma}_f^2}{n - l_f \hat{\sigma}_f^2}
\]

where \( \hat{\sigma} \) is the ML estimator of \( \sigma^2 \) in the comprehensive normal linear regression model for \( y \) given \( x \):

\[
M_{f \lor g} = (N(x', \sigma^2), \theta = (\theta', \sigma^2) \in \mathbb{R}^l \times \mathbb{R}_+).
\]

It can easily be shown that \( \Delta C_p \) and \( \Delta P \) correspond to Mallows (1973) \( C_p \) criterion and to Amemiya (1980) PC criterion respectively. For fixed explanatory variables, both \( \Delta C_p \) and \( \Delta P \) are known to be unbiased estimators of the difference in MSE risk \( E[\hat{\sigma}_g^2 - \hat{\sigma}_f^2] \), the former under the assumption that the comprehensive model \( M_{f \lor g} \) is correctly specified, and the latter under the assumption that \( M_f \) and \( M_g \) are both correctly specified (see, e.g., Judge et al. (1985, Chapter 21)). In any case, it follows from (3.7), (3.9), (3.10) and the almost sure convergence of \( \Delta \sigma^2 \) to \( \Delta \sigma^2 = \sigma_g - \sigma_f \), that \( \Delta C_p \) and \( \Delta P \) are consistent estimators of \( \Delta \text{MSE}(M_f, M_g) = \sigma_g^2 - \sigma_f^2 \) whether or not \( M_f, M_g \) and \( M_{f \lor g} \) are correctly specified. Indeed, from (3.8) - (3.10) we have:

\[
\Delta C_p = \Delta \sigma^2 + O_p(n^{-1}),
\]

\[
\Delta P = \Delta \sigma^2 + O_p(n^{-1}).
\]

Hence \( \sqrt{n} \Delta C_p \) and \( \sqrt{n} \Delta P \) will have the same asymptotic distribution as \( \sqrt{n} \Delta \sigma^2 \) whenever \( \sqrt{n} \Delta \sigma^2 \) converges in distribution to a limit. Such an approximation is useful since the exact finite sample distributions of \( \Delta \sigma^2 \), \( \Delta C_p \), and \( \Delta P \) are difficult to obtain especially when \( M_{f \lor g} \) is misspecified.\(^8\)

Next, we turn to model selection criteria based on the KLIC. From (2.27) and (3.7), a natural consistent estimator of \( \Delta \text{KLIC}(M_f, M_g) \) is:
\[ \frac{1}{n} \log \left( \frac{\sigma_g^2}{\sigma_f^2} \right) = \frac{1}{n} \left[ \sum_{i=1}^{n} f_i (\theta_f) - \sum_{i=1}^{n} f_i (\hat{\theta}_f) \right], \]

where the second equation follows from (3.6). In Vuong (1985), we derived the asymptotic distribution of the likelihood-ratio (LR) statistic under general conditions. This approximation was then used to construct some LR-based tests of the null hypothesis that the models \( M_f \) and \( M_g \) are KLIC-equivalent. It will be used in the next sections when the competing models are linear regression models.

As for the theoretical criterion based on the MSE, there exist other consistent estimators of \( \text{AKLIC}(M_f, M_g) \). In particular, we have:

\[ \frac{1}{n} \log \left( \frac{\sigma_g^2}{\sigma_f^2} \right) = \frac{1}{n} \left[ \sum_{i=1}^{n} f_i (\theta_f) - \sum_{i=1}^{n} f_i (\hat{\theta}_f) \right], \]

(3.15)

\[ \frac{1}{n} \log \left( \frac{\sigma_g^2}{\sigma_f^2} \right) = \frac{1}{n} \left[ \sum_{i=1}^{n} f_i (\theta_f) - \sum_{i=1}^{n} f_i (\hat{\theta}_f) \right], \]

(3.16)

\[ \frac{1}{n} \log \left( \frac{\sigma_g^2}{\sigma_f^2} \right) = \frac{1}{n} \left[ \sum_{i=1}^{n} f_i (\theta_f) - \sum_{i=1}^{n} f_i (\hat{\theta}_f) \right], \]

(3.17)

Criterion (3.15) corresponds to Akaike (1973) information criterion.

Criterion (3.16) corresponds to Sawa (1978) information criterion for normal linear regression models.\( ^9 \) Criterion (3.17) corresponds to Schwarz (1978) formula for discriminating between models. These criteria were derived under different assumptions. For instance, \( \text{AAIC} \) was derived under the assumption that both models \( M_f \) and \( M_g \) are correctly specified, while \( \text{ABIC} \) was derived under the assumption that the comprehensive model \( M_f \cup M_g \) is correctly specified. From (2.27), (3.6), (3.7) and (3.15) - (3.17), it is clear, however, that \( n^{-1} \text{AAIC} \), \( n^{-1} \text{ABIC} \), and \( n^{-1} \text{ASIC} \) are all strongly consistent estimators of \( \text{AKLIC}(M_f, M_g) = \frac{1}{2} \log (\sigma_g^2/\sigma_f^2) \). In addition, we have under general conditions:

\[ \frac{1}{n} \text{AAIC} = \frac{1}{n} \text{LR} \left( \hat{\theta}_f, \hat{\theta}_g \right) + o_P(n^{-1}), \]

(3.18)

\[ \frac{1}{n} \text{ABIC} = \frac{1}{n} \text{LR} \left( \hat{\theta}_f, \hat{\theta}_g \right) + o_P(n^{-1}), \]

(3.19)

\[ \frac{1}{n} \text{ASIC} = \frac{1}{n} \text{LR} \left( \hat{\theta}_f, \hat{\theta}_g \right) + o_P(n^{-1}), \]

(3.20)

for any \( \alpha > 0 \). Hence \( n^{-1/2} \text{AAIC}, n^{-1/2} \text{ABIC}, n^{-1/2} \text{ASIC} \) will have the same asymptotic distribution as \( n^{-1/2} \text{LR} \left( \hat{\theta}_f, \hat{\theta}_g \right) \) whenever \( n^{-1/2} \text{LR} \left( \hat{\theta}_f, \hat{\theta}_g \right) \) converges in distribution to a limit, as it will be the case in the next sections.

The previous discussion suggests that a classical approach for choosing between two competing models \( M_f \) and \( M_g \) is to test the null hypothesis

\[ H_0: M_f \text{ is equivalent to } M_g \]

(3.21)

against either one of the alternatives

\[ H_f: M_f \text{ is better than } M_g \]

(3.22)

\[ H_g: M_g \text{ is better than } M_f. \]

(3.23)
As shown in Section 2, the MSE criterion and the KLIC are equivalent when comparing normal linear regression models. Thus one can equivalently test $H_0$: $\text{AMSE}(M_f, M_g) = 0$ against $H_f$: $\text{AMSE}(M_f, M_g) > 0$ (or $H_o$: $\text{AKLIC}(M_f, M_g) = 0$ against $H_f$: $\text{AKLIC}(M_f, M_g) > 0$). Following Vuong (1985), we shall use the natural LR statistic $\text{LR}_{n(f, g)}$, but clearly any one of the previous statistics $\Delta \omega^2$, $\Delta C^p$, $\Delta PC$, $\text{AAIC}$, $\text{ABIC}$, and $\text{ASIC}$ can be used instead because of their small sample properties.

We shall mainly study the case where the linear regression models $M_f$ and $M_g$ are non-nested, and we shall propose some model selection tests under different information structures. Namely, we shall successively treat (i) the general case where none of the competing models is correctly specified, (ii) the case where the comprehensive model $M_f \cup M_g$ is correctly specified, and (iii) the case where at least one model is correctly specified. The classical case where the linear regression models are nested will be discussed only briefly in the conclusion.

4. THE GENERAL CASE

Let $M_f$ and $M_g$ be two normal linear regression models of $y$ given $x$, i.e., models of the form (3.1). Let $X_g$ be the vector of included explanatory variables in the model $M_g$, $s = f, g$. The models $M_f$ and $M_g$ are assumed to be non-nested. Thus, we assume:

**Assumption A5:** $X_f \notin X_g$ and $X_g \notin X_f$.

It is convenient to partition $X_f$ and $X_g$ into some common explanatory variables and some variables specific to $M_f$ and $M_g$:

$$X'_f = (x', z'), \quad X'_g = (x', w')$$

(4.1)

where $x$, $z$, and $w$ are respectively $k$, $p$, and $q$ dimensional vectors.

Thus $f_x = k + p$ and $f_g = k + q$. The coefficient vectors $\lambda_f$ and $\lambda_g$ are partitioned accordingly into:

$$\lambda_f = (a', \beta'), \quad \lambda_g = (\gamma', \delta')$$

(4.2)

Then A5 is equivalent to the assumption that $p \neq 0$ and $q \neq 0$. Without loss of generality, we shall assume throughout that $p \geq q$ and that the union of $X_f$ and $X_g$ is equal to $x$ so that $k + p + q = l$.

Strictly speaking, linear regression models can never be strictly non-nested since they must have some common conditional distributions for $y$ given $x$ (see Vuong (1985, Definition 5.1)). Indeed, it is easy to see from (3.1) that $M_f$ and $M_g$ must both contain the non-empty class of conditional distributions for $y$ given $x$:

$$M_0 = \{N(x; \lambda, \sigma^2) ; \lambda = 0, \sigma^2 \in \mathbb{R}_+\}$$

(4.3)

Hence, linear regression models can only be either overlapping (Vuong (1985, Definition 6.1)) or nested (Vuong (1985, Definition 7.1)).

The fact that $M_f \cap M_g \neq \emptyset$ even in the non-nested case has not often been recognized in the literature, and in fact much complicates the derivation of some classical tests of the null hypothesis that the models $M_f$ and $M_g$ are MSE or KLIC equivalent. Indeed, as shown in
Vuong (1985, Theorem 3.5), the asymptotic distribution of the LR statistic as well as the speed at which it converges to that distribution crucially depends on whether or not the closest distribution in \( M_f \) to the true distribution \( H_0 \) is \( H_0 \)-almost surely identical to the closest distribution in \( M \) to \( H_0 \), i.e., on whether or not

\[
H_0^*; \quad \mathcal{L}(\cdot; \theta^*_f) = \mathcal{L}(\cdot; \theta^*_g) \quad H_0 \text{-almost surely}, \tag{4.4}
\]

holds, where \( \mathcal{L}(\cdot; \theta_s) \) denotes the univariate conditional normal density of \( y \) given \( x_s \) with parameter \( \theta_s \). For \( s = f, g \), let

\[ e_s = y - x_n^s. \]

Since \( \text{var}^o(e_s) = \sigma_s^2 \), then it can readily be shown that the null hypothesis \( H_0^* \) can be equivalently rewritten as:

\[
H_0^*; \quad e_f^2 = e_g^2 \quad H_0 \text{-almost surely.} \tag{4.5}
\]

Then, applying Vuong (1985) Theorem 5.2 to the normal linear regression models we obtain:

**Proposition 4.1:** Given \( A_1 - A_5 \),

(i) under \( H_0 - H_0^* \), \( T_{fg} \rightarrow N(0, 1) \),

(ii) under \( H_f \), \( T_{fg} \rightarrow +\infty \),

(iii) under \( H_g \), \( T_{fg} \rightarrow -\infty \),

where

\[
T_{fg} = \frac{\sqrt{n} \left[ \sum_{t=1}^n \frac{\hat{\alpha}_t}{e_{ft}} - \sum_{t=1}^n \frac{\hat{\alpha}_t}{e_{gt}} \right]}{\sqrt{\sum_{t=1}^n \frac{e_{gt}^2}{e_{ft}^2} + \log \left( \sum_{t=1}^n \frac{e_{gt}^2}{e_{ft}^2} \right)}}. \tag{4.6}
\]

The statistic \( T_{fg} \) is nothing else than \( n^{-1/2} \text{LR}_n(\hat{\theta}_f, \hat{\theta}_g)/\sigma_n \), where

\[
\sigma_n = \frac{n}{N} \left[ \sum_{t=1}^n \frac{\hat{\alpha}_t}{e_{ft}} - \sum_{t=1}^n \frac{\hat{\alpha}_t}{e_{gt}} \right]^2. \tag{4.7}
\]

It is important to note that we necessarily have \( H_0^* \subset H_0 \). Thus assuming that \( H_0^* \) does not hold, Proposition 4.1 - (i) gives us a simple asymptotically normal test of the null hypothesis that the normal linear regression models \( M_f \) and \( M_g \) are equivalent. The test is directional, and Parts (ii) and (iii) ensure that the test is consistent against the alternatives \( H_f \) and \( H_g \).

As mentioned in Vuong (1985, Section 5), the statistic \( T_{fg} \) can in general be obtained from an additional linear regression. Let

\[
m_t = \frac{\hat{\alpha}_t}{e_{ft}} - \frac{\hat{\alpha}_t}{e_{gt}} + \log \left( \sum_{t=1}^n \frac{e_{gt}^2}{e_{ft}^2} \right), \]

where

\[
\eta = \frac{n}{N} \left[ \sum_{t=1}^n \frac{\hat{\alpha}_t}{e_{ft}} - \sum_{t=1}^n \frac{\hat{\alpha}_t}{e_{gt}} \right].
\]
Then, it can easily be shown that $T_{fg}$ is equal to $[n(n - 1)]^{-1/2}$ times the usual $t$-statistic on the constant term in a linear regression of $m_t$ on only the constant term.

However, the previous statistic can only be used to test $H_0 - H_{00}$. Since $H_{00}$ is part of the null hypothesis $H_0$, it is also necessary to test $H_{00}$ in order to determine whether $M_f$ and $M_g$ are equivalent. In Vuong (1985), we propose the following two-step procedure: Test $H_{00}$ against its alternative $H_{00}^*$, if $H_{00}$ cannot be rejected then the two models are equivalent, otherwise test $H_0 - H_{00}$ using the statistic $T_{fg}$ as indicated in the previous paragraph. As shown there, if $a_1$ and $a_2$ are the asymptotic significance level of these tests, then the asymptotic significance level $\alpha$ of this sequential procedure as a test of the null hypothesis of interest $H_0$ is not larger than $\max(a_1, a_2)$. Hence, if $a_1 = a_2 = 10\%$, then $\alpha \leq 10\%$.

The remainder of this section considers various ways for testing $H_{00}$. Let

$$w^2 = \text{Var}_0 \left[ \log \frac{d_f(y|x,\theta_f^*)}{d_g(y|x,\theta_g^*)} \right],$$

$$w^2 = \frac{1}{4} \left[ \frac{s_f^2}{s_f^2} - \frac{s_g^2}{s_g^2} \right]^2. \tag{4.8}$$

Recall that $\lambda_f^* = (\alpha^*, \beta^*, \gamma^*, \delta^*)$ and $\lambda_g^* = (\alpha^*, \beta^*, \gamma^*, \delta^*)$ are the pseudo-true parameters for $\lambda_f$ and $\lambda_g$ in the models $M_f$ and $M_g$ respectively. We shall also consider the comprehensive normal linear regression model for $y$ given $\chi = (x', z', w')'$ defined in (3.11). Let $\lambda^*_f = (\alpha^*_f, \beta^*_f, \gamma^*_f, \delta^*_f)$ be the pseudo-true parameters corresponding to this model. (These can be obtained from Lemma 2.1 by setting $\chi_S = \chi$.)

Then, let $e_f \mid g = y - \chi \lambda^*_f$.

**Lemma 4.2**: Given $A_2 - A_5$, the null hypothesis $H_{00}^*$ is equivalent to either one of the following statements:

1. $\omega^2 = 0$,
2. $\beta^* = 0$ and $\delta^* = 0$,
3. $\lambda^*_f = 0$ and $\lambda^*_g = 0$.

That $H_{00}^*$ is equivalent to $\omega^2 = 0$ is a general result (see Vuong (1985, Lemma 4.1)). It is easy to see that each of the latter two statements implies $H_{00}^*$. These implications do not depend on Assumption $A_2$. On the other hand, as the example in the Appendix shows, their converses crucially depend on $A_2$ which states that the true conditional distribution $H_0$ has a strictly positive density with respect to the Lebesgue measure $\nu_y$. In particular, $y$ cannot be a discrete variable or have mass points.

**Lemma 4.2** is, however, intuitively desirable. Indeed, rewriting (4.4) in the linear regression form, Part (ii) says that the conditional distributions for $y$ given $\chi = (x', z', w')'$ defined by:

$$y = x', \beta^* + e_f, \quad e_f \sim N(0, \sigma_f^2) \tag{4.9}$$

$$y = x', \gamma^* + \delta^* + e_g, \quad e_g \sim N(0, \sigma_g^2) \tag{4.10}$$
are \( H^0 \)-almost surely identical if and only if \( \beta^* = \delta^* = 0 \), or equivalently if and only if \( \lambda^*_z = \lambda^*_w = 0 \) in the comprehensive conditional distribution for \( y \) given \( (x', z', w')' \):

\[
y = x' \lambda^*_x + z' \lambda^*_z + w' \lambda^*_w + e_f \sigma_f \Gamma_f + e_g \sigma_g \Gamma_g \sim N(0, \sigma_f^2 \Gamma_f \sigma_g^2 \Gamma_g).
\]  

(4.11)

If either one of this conditional holds, then

\[
a^* = \gamma^* = \lambda^*_x
\]  

(4.12)

\[
\sigma_f^2 = \sigma_g^2 = \sigma_f^2 \Gamma_f \sigma_g^2 \Gamma_g
\]  

(4.13)

\[
e_f = e_g = e_f \sigma_f \Gamma_f
e_g \sigma_g \Gamma_g
\]  

(4.14)

(see the proof of Lemma 4.2).

Lemma 4.2 allows us to test \( H^0 \) in various ways. For instance, using Part (i), we can test \( H^0 \) by using the estimator \( \tilde{\omega}_n \) defined in (4.7). This is the general procedure proposed in Vuong (1985, Theorem 4.4) where it is shown that the statistic

\[
\tilde{\omega}_n
\]  

is asymptotically distributed under \( H^0 \) as a weighted sum of chi-squares with weights equal to the squares of the eigenvalues of the matrix

\[
W = \begin{bmatrix}
-\frac{1}{\sigma_f^2 \Gamma_f} & -\frac{1}{\sigma_g^2 \Gamma_g} \\
\frac{1}{\sigma_f^2 \Gamma_f} & \frac{1}{\sigma_g^2 \Gamma_g}
\end{bmatrix}
\]  

(4.15)

where, as usual,

\[
A_s = \mathbb{E}_s \left[ \sigma^2 \log L_s \left( y | x_s ', \theta_s \right) \right] \frac{\delta \log L_s \left( y | x_s ', \theta_s \right)}{\delta \theta_s} \frac{\delta \log L_s \left( y | x_s ', \theta_s \right)}{\delta \theta_s'}
\]  

(4.16)

\[
B_s = \mathbb{E}_s \left[ \sigma^2 \log L_s \left( y | x_s ', \theta_s \right) \right] \frac{\delta \log L_s \left( y | x_s ', \theta_s \right)}{\delta \theta_s} \frac{\delta \log L_s \left( y | x_s ', \theta_s \right)}{\delta \theta_s'}
\]  

(4.17)

\[
B_{fg} = \mathbb{E}_s \left[ \sigma^2 \log L_s \left( y | x_s ', \theta_s \right) \right] \frac{\delta \log L_s \left( y | x_s ', \theta_s \right)}{\delta \theta_f} \frac{\delta \log L_s \left( y | x_s ', \theta_s \right)}{\delta \theta_g'}
\]  

(4.18)

with \( \Theta_s = (\lambda^*_s, \sigma_s^2)' \) and \( s = f, g \). The next lemma determines these matrices under the hypothesis \( H^0 \). Given Lemma 4.2, we can define:

\[
e = e_f = e_g = e_f \sigma_f \Gamma_f
\]  

(4.19)

\[
\sigma_f^2 = \sigma_g^2 = \sigma_f^2 \Gamma_f \sigma_g^2 \Gamma_g
\]  

(4.20)

Lemma 4.3: Given \( A_2 - A_5 \), under \( H^0 \)

\[
A_s = \begin{bmatrix}
\sum_{\theta \in \Theta_s} \sigma^2 \log L_s \left( y | x_s ', \theta \right) & 0 \\
0 & 1/(2 \sigma^2)
\end{bmatrix}
\]  

(4.21)

\[
A_s = \begin{bmatrix}
\sigma^2 \log L_s \left( y | x_s ', \theta_s \right) & 0 \\
0 & 1/(2 \sigma^2)
\end{bmatrix}
\]  

(4.22)

\[
A_s = \begin{bmatrix}
\sigma^2 \log L_s \left( y | x_s ', \theta_s \right) & 0 \\
0 & 1/(2 \sigma^2)
\end{bmatrix}
\]  

(4.23)
In the general case where the models \(M_f\) and \(M_g\) are not necessarily correctly specified, the matrix \(W\) does not simplify since the information matrix equivalence \(A_s + B_s = 0\) does not necessarily hold (see, e.g., White (1982))\(^{12}\). Determination of the eigenvalues of \(W\) is, however, important for the subsequent tests. Since \(W\) is of dimension \(l_f + l_g + 2 = 2k + p + q + 2\), the matrix \(W\) has 2k + p + q + 2 eigenvalues, which are all real (see Vuong (1985)).

The next lemma is quite useful since it states that at least 2k + 2 eigenvalues are zero, and shows how the remaining eigenvalues can be obtained. We need some additional notation. Define

\[
R' = \begin{bmatrix}
-\Sigma_{xx}^{-1} & \Sigma_{xx}^{-1} & 0 \\
-\Sigma_{xz}^{-1} & \Sigma_{xz}^{-1} & 0 \\
0 & 0 & I_p
\end{bmatrix},
\]

\[
C = \begin{bmatrix}
C_{xx} & C_{xz} & C_{xw} \\
C_{zx} & C_{zz} & C_{zw} \\
C_{wx} & C_{zw} & C_{ww}
\end{bmatrix},
\]

\[
Q = \begin{bmatrix}
Q_{zz} & Q_{zw} \\
Q_{zw} & Q_{ww}
\end{bmatrix},
\]

where

\[
C_{ij} = \text{Cov}^0(\epsilon_i, \epsilon_j),
\]

\[
Q_{ij} = \sum_{l=1}^{l_f} l^2 - \sum_{l=1}^{l_f} l \sum_{x}^{O} \sum_{y}^{O} -1 \sum_{x}^{O} \sum_{j}^{O}.
\]

for \(i, j = x, z, w\). Define also

\[
\text{Diag}(Q_{zz} - Q_{ww}) = \begin{bmatrix}
Q_{zz} & 0 \\
0 & -Q_{ww}
\end{bmatrix}.
\]

It can easily be shown that under \(A_2\), \(Q\) and hence \(\text{Diag}(Q_{zz} - Q_{ww})\) are both non-singular.

**Lemma 4.4:** Given Assumptions \(A2 - A5\), under \(H_0^*\), the matrix \(W\) has at least 2k + 2 zero eigenvalues, the other p+q eigenvalues \(\lambda_u\) are real and are solutions of

\[
\det[R'CR - \lambda_u^2 \text{Diag}(Q_{zz} - Q_{ww})] = 0.
\]

With all the eigenvalues of \(W\) being characterized, the test of \(H_0^*\) against \(H_A^*\) (that is, the variance test in terms of Vuong (1985)) using \(n\) as the statistic has the following property:

**Proposition 4.5 (Variance Test):** Given Assumptions \(A1 - A5\),

1. under \(H_0^*\), \(n\) a.s. \(\rightarrow M_{p+q}(\cdot; \lambda_u^2)
2. under \(H_A^*\), \(n\) a.s. \(\rightarrow +\infty.

where \(M_{p+q}(\cdot; \lambda_u^2)\) is a weighted sum of chi-squares with weights equal to \(\lambda_u^2\).
In practice, the eigenvalues $\lambda_w$ are unknown and must be consistently estimated. This can clearly be done by consistently estimating the unknown matrices in (4.30) by their sample analogs:

\[ \hat{C}_{ij} = \frac{1}{n} \sum_{t=1}^{n} x_t' x_t \]

(4.31)

\[ \hat{f}_{ij} = \sum_{t=1}^{n} x_t' x_t \]

(4.32)

\[ \hat{C}_{ij} = \frac{1}{n} \sum_{t=1}^{n} x_t' x_t \]

(4.33)

where $e_t$ can be taken to be $y_t - x_t' \hat{\beta}, y_t - x_t' \hat{\beta}, y_t - x_t' \hat{\beta},$ or more directly, $y_t - x_t' \hat{\beta}, x_t' \hat{\beta}, y_t - x_t' \hat{\beta}, x_t' \hat{\beta}, y_t - x_t' \hat{\beta}, x_t' \hat{\beta}, y_t - x_t' \hat{\beta}, x_t' \hat{\beta}, y_t - x_t' \hat{\beta}, x_t' \hat{\beta}, w_t' w$

because of Lemma 4.2 and (4.19).

Estimation of the non-zero eigenvalues $\lambda_w$ can be avoided if one knows a priori that all the non-zero eigenvalues are equal to one. As seen in the next section, this will not be the case in general. Hence the variance statistic is not necessarily chi-square distributed under $H_{o.w}$. On the other hand the test is consistent against all alternatives.

As indicated in Lemma 4.2, $H_{o.w}$ is also equivalent to $\beta^* = \delta^* = 0$. Thus, a Wald test based on an appropriate quadratic form in $(\hat{\beta}', \hat{\delta}')$ may replace the variance test. The next lemma gives the asymptotic covariance matrix of $n^{1/2}(\hat{\beta}', \hat{\delta}')$ under the null hypothesis $H_{o.w}$. Let $\sum_{xy} = \left( \sum_{xy}, \sum_{zy}, \sum_{wy} \right)'$ where a quantity with a hat denotes the sample analog of that quantity. For instance $\sum_{xy} = \frac{1}{n} \sum_{t=1}^{n} x_t' y_t$.

Lemma 4.6: Given Assumptions A1 - A5,

(1) $(\hat{\beta}', \hat{\delta}') \sim \text{Diag}(Q^{-1}, Q^{-1}) R \sum_{xy}$

(4.34)

(2) under $H_{o.w}$, $n^{1/2}(\hat{\beta}', \hat{\delta}') \overset{D}{\sim} N(0, V)$ where

\[ V = \text{Diag}(Q^{-1}, Q^{-1}) R \sum_{xy} \text{Diag}(Q^{-1}, Q^{-1}), \]

(4.35)

Since $R$ has full-column rank (see Equation (4.24)), it follows that if $C$ is nonsingular, then the asymptotic covariance matrix $V$ is non-singular. In general, however, $V$ will be singular. Thus, using generalized (g-) inverses, we define a Wald statistic as:

\[ W_n = n(\hat{\beta}, \hat{\delta}) G_n(\hat{\beta}, \hat{\delta}) \]

a.s.

(4.36)

where $G_n \rightarrow G$ and $G$ is a g-inverse of $V$, i.e., $G = V^{-1}$ (see Moore (1977), Vuong (1986)). Let $r = \text{rank } V \leq p + q$ and let

\[ \mathcal{H}_{A}^{(w)} = \{(\beta', \delta')' \in M(V) - \{0\}\}

(4.37)

where $M(V)$ denotes the $r$-dimensional manifold generated by the columns of $V$. Note that $\mathcal{H}_{A}^{(w)}$ is equivalent to $(\beta', \delta')' \neq 0$ (see Lemma 4.2). We have:

Proposition 4.7: Given Assumptions A1 - A5,
under $H^p_0$, for any choice of $g$-inverse $G$ and for any consistent
sequence $G_{n}^{'}, W_{n}^1 \rightarrow \chi^2_r$.

(ii) under $H^p_A$, $W_{n}^1 \overset{a.s.}{\rightarrow} \infty$.

Contrary to the variance statistic, the Wald statistic $W_{n}^1$ is
always chi-square distributed under $H^p_0$. Hence the test based on $W_{n}^1$
is easier to carry out. On the other hand, if $V$ is singular so that
$r < p + q$, the Wald test would not be consistent against all
alternatives in $H^p_A$. Indeed, there would exist a $p + q - r$ dimensional
manifold of alternatives against which the Wald test will not have any
asymptotic power (see Vuong (1986)).

The previous Wald test requires two OLS regressions to obtain
$\hat{\beta}$ and $\hat{\gamma}$. Using Lemma 4.2 (iii), we can think of testing $H^p_0$
by testing instead $\lambda^*_z = \lambda^*_w = 0$ in the comprehensive normal linear
regression model $M_{fVg'}$, and hence to do only one OLS regression. Let
$\hat{\lambda}_z$ and $\hat{\lambda}_w$ be the OLS estimates of $\lambda^*_z$ and $\lambda^*_w$ for this comprehensive
model.

Lemma 4.8: Given Assumptions A1 - A5,

(i) $\begin{pmatrix} \hat{\lambda}^*_z \\ \hat{\lambda}^*_w \end{pmatrix} \overset{D}{\rightarrow} \Sigma_Y^{-1} \Sigma_{XY}$

(ii) under $H^p_0$, $n^{1/2} \begin{pmatrix} \hat{\lambda}^*_z \\ \hat{\lambda}^*_w \end{pmatrix} \overset{D}{\rightarrow} N(0, W)$ where

$W = Q^{-1} R' CRQ^{-1}$. (4.39)

Note that rank $W = \text{rank} V = r$. As before, we define a Wald
statistic as:

$w^2 = n(\hat{\lambda}^*_z, \hat{\lambda}^*_w)' H_n (\hat{\lambda}^*_z, \hat{\lambda}^*_w)'$ (4.40)

where $H_n \rightarrow H$ and $H$ is a $g$-inverse of $W$. The next result relates $w^2$
to $W_{n}^1$ and gives the asymptotic properties of the Wald test based on
$w^2$. Let

$H^w_A = \{(p^*, q^*)' \in M(W) - \{0\}\}$ (4.41)

Proposition 4.9: Given Assumptions A1 - A5,

(i) under $H^p_0$, for any choice of $G$, $H$, and consistent sequences $G_{n}$
and $H_{n}, W_{n}^1 \rightarrow \chi^2_r$.

(ii) under $H^p_A$, for any choice of $H$ and consistent sequence $H_{n}, W_{n}^1 \rightarrow \chi^2_r$.

(iii) under $H^w_A$, $w^2 \rightarrow \infty$.

Thus the second Wald test has the same asymptotic properties
as the first Wald test. It is asymptotically chi-square distributed
under $H^p_0$, but it is not consistent against all alternatives in $H^p_A$ if
$r < p + q$.

5. THE COMPREHENSIVE MODEL IS CORRECTLY SPECIFIED

A well-known and important method for discriminating between
two competing non-nested models is to construct a so-called
comprehensive model that contains both competing models. This
approach was initially suggested by Cox (1961, 1962) and subsequently studied by Atkinson (1970). When the exponential combination of the competing densities is used, we obtain for the normal linear regression models $M_f$ and $M_g$:

$$y = x'[(1 - \lambda)\frac{2, \alpha}{\sigma_f^2} + \lambda\frac{2, \gamma}{\sigma_g^2}] + (1 - \lambda)\frac{2, \delta}{\sigma_f^2} + \lambda\frac{2, \delta}{\sigma_g^2} + u$$

where $u \sim N(0, \sigma^2)$, and $\sigma^2 = (1 - \lambda)\sigma_f^2 + \lambda\sigma_g^2$ (see, e.g., Pesaran (1982)). It can readily be shown that the model (5.1) for $y$ given $x$ is identical to the model $M_{f'g}$ defined in Equation (3.11).

When $\lambda = 0$, the model (5.1) reduces to $M_f$, and when $\lambda = 1$, it reduces to $M_g$. Then, assuming that the comprehensive model (5.1) [or $M_{f'g}$] is correctly specified, one successively tests the hypotheses $\lambda = 0$ and $\lambda = 1$ to determine which of the two models $M_f$ and $M_g$ is best. The comprehensive approach suffers, however, from (i) the arbitrariness in the choice of a comprehensive model, (ii) the necessity of carrying out two successive tests, (iii) the fact that all the parameters $\alpha, \beta, \gamma, \delta, \sigma_f^2, \sigma_g^2,$ and $\lambda$ are not identified, and (iv) that under $\lambda = 0$ (or $\lambda = 1$) the parameters $\delta$ (or $\beta$) do not enter into the combined density so that the LR test or LM test of $\lambda = 0$ (or $\lambda = 1$) is not applicable (see, e.g., Pesaran (1982)).

In this section, we shall retain the assumption that the comprehensive model $M_{f'g}$ is correctly specified, and we shall simplify the general model selection procedure of the previous section given this additional assumption. It is clear that we can still use the simple directional normal statistic $T_{fg}$ discussed in Proposition 4.1 in order to test part of the null hypothesis $H_0$, namely $H_0 - H_0^{W}$, against $H_f$ or $H_g$. To test the remainder of the null hypothesis $H_0$, namely $H_0^{W}$, we can consider the variance and the Wald tests discussed earlier in Propositions 4.5, 4.7, and 4.9. The purpose of this section is to simplify these latter tests when the comprehensive model $M_{f'g}$ is assumed to be correctly specified.

As a matter of fact, we shall only assume that the true conditional mean of $y$ give $x$ is linear in $x$ and that the conditional variance is independent of $x$. Formally we assume

**Assumption A6:**

(a) $E(y|x) = \lambda_0 + x^T\lambda_0 + x^T\lambda^* + w^T\lambda^*$; (b) $\text{var}(y|x) = \sigma^2_0$.

It is clear that A6 is weaker than the assumption that the comprehensive model $M_{f'g}$ is correctly specified. Therefore our results naturally apply to the latter case. The following lemma presents the implications of $H_0$ on the true conditional mean and variance when A6 is satisfied.

**Lemma 5.1:** Given Assumptions A2 - A6, under $H_0$, we have

(i) $\lambda_0 = \lambda^* = \beta^* = 0$; $\lambda_0^{W} = \lambda^* = \delta^* = 0$,

(ii) $\lambda_0^{X} = \lambda_0 = \lambda^* = \gamma^*$,

(iii) $\sigma^2_0 = \sigma^2_{f'g} = \sigma^2_f = \sigma^2_g$. 
That is, the imposition of Assumption A6 ensures that the true conditional expectation of \( y \) only depends on \( x \) under \( H_0^\circ \), consequently all the pseudo-true parameters are equivalent to their corresponding true parameters. From Lemma 5.1, it is clear that the addition of A6 will simplify the general expressions. In particular, under \( H_0^\circ \) for any \( i,j = x, z, w, \)
\[
C_{ij} = E(i_j e) = E(i_j E(e|x,z,w)) = c^2 \alpha_{ij},
\]
where \( e = y - x a = y - x x' = y - x x' \) so that
\[
C = c^2 \alpha.
\]

Hence \( C \) is non-singular under \( H_0^\circ \).

In addition the matrices \( Q_{zz}, Q_{ww}, \) and \( Q_{zw} \) have a natural interpretation. For example, \( Q_{ww} = \text{Var}^0(e_1) \) where \( e_1 = w - a x \) and \( a^* = \sum_{xx} \left( \sum_{xx} \right)^{-1}; Q_{zz} = \text{Var}^0(e_2) \) where \( e_2 = z - b x \) and \( b^* = \sum_{xx} \left( \sum_{xx} \right)^{-1} \). That is, we artificially set up two linear regression models with \( w \) and \( z \) being the two dependent variables, \( x \) being the common independent variable. Therefore, \( Q_{ww} \) is the variance of the residual for the first model while \( Q_{zz} \) is the variance of the residual for the second model. Moreover, if we artificially set up a linear regression model with \( e_1 \) being the dependent variable, \( e_2 \) being the independent variable, then \( \text{Var}^0(e_3) = Q_{ww} - Q_{zw}^2 Q_{zw}^{-1} \) where \( e_3 = e_1 - c e_2 \) and \( c^* = Q_{zw}^{-1} \).

**Lemma 5.2:** Given A2 - A6, under \( H_0^\circ \), \( W \) has exactly \((2k + 2)\) zero eigenvalues, \( p\)-rank \( Q_{zw} \) eigenvalues equal to one, and \( q\)-rank \( Q_{zw} \) eigenvalues equal to minus one. The remaining \( 2 \) rank \( Q_{zw} \) eigenvalues \( \lambda_{wi} \), if any, are real with \( 0 < \left| \lambda_{wi} \right| < 1 \), and solve:
\[
|Q_{zw}^2 Q_{zw}^{-1} - (1 - \lambda_{wi}^2)|_2 = 0.
\]

Putting \( \mu = (1 - \lambda_{wi}^2) \), Equation (5.4) can be solved by determining the eigenvalues of \( Q_{ww}^{-1/2} Q_{zw} Q^{-1/2} \) or \( Q_{ww}^{-1} Q_{ww}^{-1} \). This latter matrix has an interesting interpretation in terms of the vectors of random variables \( e_1 \) and \( e_2 \) defined above. Indeed, we can write \( Q_{ww}^{-1} Q_{zw} Q_{zw}^{-1} \) as \([\text{Var}^0(e_1) - \text{Var}^0(e_2)]\text{Var}^0(e_1)^{-1} \), which can be treated as a generalized version of \( R^2 \) for the artificial linear regression model with \( e_1 \) being the dependent variable and \( e_2 \) being the independent variable. A particular case is when \( p = q = 1 \) so that \( Q_{ww}^{-1} Q_{ww}^{-1} \) is the usual \( R^2 \) for the artificial population regression of \( e_1 \) on \( e_2 \).

From Proposition 4.5 and Lemma 5.2, we obtain:

**Proposition 5.2:** Given Assumptions A1 - A6,

(i) under \( H_0^\circ \), \( n^{\omega \circ} \rightarrow \chi^2_{p+q-2} \) rank \( Q_{zw} \) + \( \sum_{i=1}^{\text{rank} Q_{zw}} \lambda_{wi}^2 \chi^2(2), \)

where \( 0 < \lambda_{wi} < 1 \).

(ii) under \( H_0^\circ \), \( n^\circ \rightarrow +\infty \).
Corollary 5.4: Given Assumptions A1 - A6, under \( H_0^w \), \( \hat{\sigma}^2_n \) is asymptotically chi-square distributed if and only if \( Q_{zw} = 0 \).

Since \( Q_{zw} = \sum_{zw}^0 - \sum_{zw}^0 (\sum_{zw}^0)^{-1} \sum_{zw}^0 \), the condition \( Q_{zw} = 0 \) can be interpreted as requiring that the variables \( z \) and \( w \) are conditionally orthogonal given \( x \). This is satisfied, for instance, if \( z \) is orthogonal to \((x, w)\) so that \( \sum_{zx}^0 = \sum_{zw}^0 = 0 \), or if \( w \) is orthogonal to \((x, z)\) so that \( \sum_{wx}^0 = \sum_{zw}^0 = 0 \). If there are no common explanatory variables \( x \), then \( Q_{zw} = 0 \) is equivalent to \( z \) and \( w \) being orthogonal, i.e., \( \sum_{zw}^0 = 0 \).

Although the variance test generally involves the distribution of a weighted sum of chi-squares, following Section 4 we may also employ Wald statistics to test \( H_0^w \) against \( H_A^w \) where asymptotic chi-square distributions prevail. As noticed earlier, under \( H_0^w \), the covariance matrix \( C \) is non-singular so that the use of g-inverses in Equations (4.36) and (4.40) become unnecessary. Hence we can define the Wald statistics based on \((\hat{\beta}', \hat{\delta}')'\) and \((\hat{\gamma}_z', \hat{\lambda}_w')\) directly as:

\[
W_n^1 = n(\hat{\beta}', \hat{\delta}')V_n^{-1}(\hat{\beta}', \hat{\delta}')',
\]

\[
W_n^2 = n(\hat{\gamma}_z', \hat{\lambda}_w')W_n^{-1}(\hat{\gamma}_z', \hat{\lambda}_w')',
\]

where \( V_n \) and \( W_n \) are consistent estimators under \( H_0^w \) of the non-singular asymptotic covariance matrices \( V \) and \( W \). From (4.35) and (4.39), we can clearly choose the consistent estimates obtained by replacing in these formulae the matrices \( Q \) and \( R \) by their sample analogs \( \hat{Q} \) and \( \hat{R} \), and the matrix \( C \) by

\[
\hat{C} = \hat{\sigma}^2_n \hat{C},
\]

\[
\hat{\sigma}^2 = \frac{1}{n} \sum_{t=1}^{n} \epsilon_t^2,
\]

where \( \epsilon_t \) can be taken to be \( y_t - x_t' \hat{\beta}, y_t - x_t' \hat{\delta}, y_t - x_t' \hat{\gamma}_x, y_t - x_t' \hat{\lambda}_x - z_t' \hat{\beta}, y_t - x_t' \hat{\gamma}_z - z_t' \hat{\lambda}_z - w_t' \hat{\lambda}_w \). For any choice of \( \epsilon_t \), we have:

Proposition 5.5: Given Assumptions A1 - A6,

(i) \( W_n^1 \) and \( W_n^2 \) converge in distribution to \( \chi^2_{p+q} \),

(ii) under \( H_0^w \), any of these statistics converges almost surely to \( +\infty \).

Thus, contrary to the variance statistic, the Wald statistics are always asymptotically chi-square distributed. Moreover, unlike the general case (see Section 4), the Wald test are consistent against all alternatives in \( H_A^w \). Thus the Wald tests are preferable to the variance test since they are easier to carry out. Finally, let us note that A2 and A6 are automatically satisfied if the comprehensive model \( M_{fVg} \) is correctly specified, i.e., if we assume:
Assumption A6*: \( H^0 = \mathcal{N}(\lambda^0, \sigma^0). \)

In this case, if \( \hat{\sigma}^2 \) is defined by (5.8) using the OLS residuals for the comprehensive model \( M_{fVg} \), i.e., \( e_t = y_t - x_t \hat{\lambda}_x - z_t \hat{\lambda}_z - \hat{w}_t \), then we can in fact obtain, under \( H_0^w \), the exact small sample distribution of the Wald statistics \( \hat{W}^1_n \) and \( \hat{W}^2_n \) conditional on \( \tilde{X} \).

Indeed, from the theory of linear regression models, we have under \( H_0^w \) and given \( \tilde{X} \):

\[
\frac{n - (k + p + q)}{n(p + q)} \hat{W}^1_n = \frac{n - (k + p + q)}{n(p + q)} \hat{W}^2_n \sim F(p + q, n - (k + p + q)).
\]

6. AT LEAST ONE MODEL IS CORRECTLY SPECIFIED

In this section, we assume that at least one of the two competing models is correctly specified, i.e.,

Assumption A7: \( H^0 \in M_f \cup M_g. \)

First, it is worth noting that A7 is stronger than the assumption considered in the previous section that the comprehensive model \( M_{fVg} \) is correctly specified. This follows from the fact that \( M_f \cup M_g \) is included in \( M_{fVg} \). Second, assuming that one knows that one model is correctly specified does not mean that one knows which one is the correct model. Indeed, if this was the case, the correct model would be at least as good as the other model (see Section 2) and the model selection problem would be trivial. Third, though A7 appears to be more rhetorical than justified in practice, such an assumption is often considered in the model selection literature where one chooses a model in a list of competing models, one of which being correct.

As noticed in Vuong (1985, Lemma 6.3), the major difficulty arising from the discrepancy between \( H_0^w \) and \( H_0^0 \) disappears since \( H_0^w = H_0^0 \) when at least one model is correctly specified. It follows that the speed at which the LR statistic converges to a limiting distribution remains constant under the null hypothesis \( H_0^w \).

Specifically, Theorem 6.4 in Vuong (1985) establishes that (i) under \( H_0^w \), \( 2LR_n(\hat{\theta}_f, \hat{\theta}_g) \) converges in distribution to a weighted sum of chi-squares with weights equal to the eigenvalues of \( W \), (ii) under \( H_f \), \( 2LR_n(\hat{\theta}_f, \hat{\theta}_g) \to +\infty \), and (iii) under \( H_g \), \( 2LR_n(\hat{\theta}_f, \hat{\theta}_g) \to -\infty \). Hence, in this case, we can bypass the sequential procedure of Sections 4 and 5 which is based on the variance (or Wald) tests followed by the normal LR test, and use directly the statistic \( 2LR_n(\hat{\theta}_f, \hat{\theta}_g) \) to choose between \( M_f \) and \( M_g \). When the models are normal linear regressions, we have from (3.6):

\[
2LR_n(\hat{\theta}_f, \hat{\theta}_g) = -(n/2) \log (\hat{\sigma}_f^2/\hat{\sigma}_g^2). \tag{6.1}
\]

In addition, a simplification in \( W \) arises since under \( H_0^w \), which is equal to \( H_0^0 \), both models must be correctly specified. Hence the usual information matrix equivalence \( A_S + B_S = 0 \) holds so that (4.15) becomes

\[
W = \begin{bmatrix}
I & B_f & B_g \\
-B_f & I & B_g \\
B_g & B_f & I
\end{bmatrix} \tag{6.2}
\]
Moreover since $H_0 = H_0^w$, and since $A7$ clearly implies $A6$, the eigenvalues of $W$ under $H_0$ can be directly obtained from Lemma 5.2. Proposition 6.1 follows.

**Proposition 6.1:** Given $A1 - A5$, and $A7$,

(i) under $H_0$,

$$2\text{LR}_{n}(\hat{\Theta}_f, \hat{\Theta}_g) \overset{D}{\to} \chi^2_{p\text{-rank}Qzw} - \chi^2_{q\text{-rank}Qzw} + \sum_{i=1}^{\text{rank}Qzw} \lambda_{w_i}(X(1) - X(1))$$

where $0 < \lambda_{w_i} < 1$, and $\lambda_{w_i}$ solves equation (5.4).

(ii) under $H_f$, $2\text{LR}_{n}(\hat{\Theta}_f, \hat{\Theta}_g) \to +\infty$.

(iii) under $H_g$, $2\text{LR}_{n}(\hat{\Theta}_f, \hat{\Theta}_g) \to -\infty$.

As in Corollary 5.4, an interesting case is when $Qzw = 0$.

**Corollary 6.2:** Given $A1 - A5$, and $A7$, if $Qzw = 0$, then under $H_0$:

$$2\text{LR}_{n}(\hat{\Theta}_f, \hat{\Theta}_g) \overset{D}{\to} \chi^2_{p\text{-rank}(Qzw)} - \chi^2_{q\text{-rank}(Qzw)}.$$  

(6.3)

That is, $2\text{LR}_{n}(\hat{\Theta}_f, \hat{\Theta}_g)$ has an asymptotic distribution which can be decomposed as the difference between two chi-squares with degrees of freedom being $p\text{-rank}(Qzw)$ and $q\text{-rank}(Qzw)$, respectively. But, whether or not $Qzw = 0$, it is worth noting that $2\text{LR}_{n}(\hat{\Theta}_f, \hat{\Theta}_g)$ can never have an asymptotic chi-square distribution. Proposition 6.1 shows that the proposed LR-based test is directional and consistent, and hence can be directly use to choose between $M_f$ and $M_g$ when at least one model is known to be correctly specified.

Finally, let us also note that, since $H_0 = H_0^w$, one can think of testing $H_0$ by applying the variance and Wald tests studied in the previous section. These latter tests are, however, not directional so that, in case of rejection of $H_0$, one cannot infer which of the two competing models is best.

### 7. CONCLUSION

In this paper, we propose a classical hypothesis approach for choosing between two normal linear regression models which may be both incorrectly specified. This approach is based on testing the null hypothesis $H_0$ that the models are MSE or KLIC equivalent against the hypothesis that one model is closer to the truth. In general the procedure is sequential and consists in testing the stronger hypothesis $H_0^w$ that the closest distributions to the truth in the competing models are identical, and in case of rejection of $H_0^w$ to test the hypothesis $H_0 - H_0^w$. The asymptotic significance level of the procedure as a test of $H_0$ is however not larger than the maximum of the chosen asymptotic significance level of each test.

To test $H_0 - H_0^w$, we propose a very simple directional and symmetric test based on the LR-statistic appropriately normalized which is asymptotically standard normal distributed under $H_0 - H_0^w$. To test $H_0^w$, we propose three tests based on the so-called variance
statistic and two Wald statistics that use either the coefficient estimates of both models or the coefficient estimates of a comprehensive linear regression model. The relationship and the consistency of these tests are studied. When the comprehensive linear model is correctly specified, the Wald based tests are identical and consistent against all alternatives to $H^0_0$. In addition, implementation of the variance test simplifies. In particular, it becomes a chi-square test when the specific variables in the competing models are conditionally orthogonal given the common variables.

An important case where one does not need to use the above sequential procedure is when one competing model is known to be correctly specified. In this case $H_0 = H^0_0$, and we propose a directional and symmetric test of $H_0$ based directly on twice the LR-statistic which is distributed as a weighted sum of chi-squares under the null hypotheses $H_0$ that the competing models are equivalent.

The previous sections were solely concerned with non-nested models. When the competing models are nested, the classical solution to the model selection problem within the hypothesis testing framework. Specifically, for the competing nested normal linear regression models:

$$M_f: y = x' a + z' \beta + e_f, \quad e_f \sim N(0, \sigma^2_f).$$

$$M_g: y = x' r + e_g, \quad e_g \sim N(0, \sigma^2_g),$$

the classical solution consists in testing $H^0_0$: $\beta = 0$ against $H^0_A$: $\beta^* \neq 0$. As Vuong (1985, Lemma 7.2) showed, the classical hypotheses $H^0_0$ and $H^0_A$ are in fact equivalent to the model selection hypotheses $H_0$ and $H_f$, respectively. Thus, our model selection approach has the desirable property that it coincides with the usual classical hypothesis approach when the competing models are nested. In other words, we can view our model selection approach as extending the classical nested hypothesis testing to non-nested situations.

As mentioned in the introduction, another solution which has recently attracted a lot of attention derives from Cox (1960, 1961)'s work on testing non-nested hypotheses. When the competing models are nested, Cox's approach does not however coincide with the classical hypothesis approach in the sense that the implicit null hypothesis of the Cox test is not identical to the usual classical parametric null hypothesis. When the competing models are non-nested, Cox tests have been used to select among models (see, e.g., Pesaran and Deaton (1978)). The procedure is based on two successive tests designed to test the validity of each competing model. For instance, when the competing models are normal linear regression models, Pesaran (1974) showed that the numerator of the Cox-statistic for testing the validity of $M_f$ is, in our notation, proportional to:

$$\log \frac{\sigma^2_f}{\sigma^2_f + \frac{1}{n_f} X_{f}^T M_f X_f}$$

where $M_g = I - X_g (X_g' X_g)^{-1} X_g'$, and $Q_{zw} \neq 0$. It is well-known that there are nine possible outcomes to this procedure, three of which are asymptotically impossible (see, e.g., Dastoor (1981)). Vuong (1985)
has, however, shown that three of the remaining outcomes are
indecisive in the sense that one cannot infer if one model is
(strictly) better than the other. This is expected since Cox tests
were initially proposed as diagnostic (or model specification) tests
and not as model selection tests. On the other hand, our proposed
model selection tests can be also thought as diagnostic tests. Indeed
if the equivalence between $M_f$ and $M_g$ is rejected in favor of $M_g$ being
better, then $M_f$ must be incorrectly specified (even though the better
model $M_g$ may still be incorrectly specified).

APPENDIX

Proof of Lemma 2.1: Since $(\lambda^*_C, \lambda^*_g, \sigma^*_g)$ and $(\lambda^*_C, \lambda^*_g, \sigma^*_g)$ maximize (2.16) and
$-E^0(y - X_g \beta_g - \lambda^*_g)^2$ respectively, the results follow immediately from
the first order conditions.

Proof of Corollary 2.2: See Johnson and Kotz (1972, p.70).

Proof of Proposition 2.3: (i) Since $\lambda^+_C = \lambda^*_C, \lambda^+_g = \lambda^*_g$, we have
$E^0(y - X_g \beta_g - \lambda^*_g)^2 = E^0(y - X_g \beta_g - \lambda^*_g)^2 = \sigma^*_g$. Therefore
AMSE($M_f, M_g$) = $\sigma^*_f - \sigma^*_g$. (ii) Upon substituting $\lambda^*_C, \lambda^*_g$, and $\sigma^*_g$ into
(2.10) we have:

$$E^0[\log d_g(y|x_g; \theta^*_g)] = \frac{1}{2} \log \sigma^*_g - \frac{1}{2} - \frac{1}{2} \log 2\pi.$$  \hspace{1cm} (A.1)

Consequently $\Delta K L I C$($M_f, M_g$) = $\frac{1}{2} \log (\sigma^*_g / \sigma^*_f)$.

Proof of Proposition 4.1: From (3.6) and (3.14), we have:

$$\text{LR}_n(\hat{\theta}_f, \hat{\theta}_g) = \sum_{t=1}^{n} \log [d_f(y_t|z_{ft}; \hat{\theta}_f) / d_g(y_t|z_{gt}; \hat{\theta}_g)]$$

$$= -\frac{n}{2} \log (\sigma^2_{g_t} / \sigma^2_{f_t}) = -\frac{n}{2} \log (\sum_{t=1}^{n} \sigma^2_{g_t} / \sum_{t=1}^{n} \sigma^2_{f_t}).$$  \hspace{1cm} (A.2)

Moreover,

$$[\log [d_f(y_t|z_{ft}; \hat{\theta}_f) / d_g(y_t|z_{gt}; \hat{\theta}_g)]^2$$
\[
\begin{align*}
&\frac{1}{2} \log \left( \frac{a_0^2}{a'_0^2} \right) + \frac{1}{2} \left( \frac{a_0^2}{a'_0^2} - \frac{a_t^2}{a'_t^2} \right)^2 \\
&= \frac{1}{4} \left[ \left( \frac{a_0^2}{a'_0^2} \right)^2 + \frac{1}{2} \log \left( \frac{a_0^2}{a'_0^2} \right) \right] + \frac{1}{2} \frac{a_t^2}{a'_t^2} - \frac{a_t^2}{a'_t^2} \\
&= \frac{1}{4} \left[ \left( \frac{a_0^2}{a'_0^2} \right)^2 + \frac{1}{2} \log \left( \frac{a_0^2}{a'_0^2} \right) \right] + \frac{1}{2} \left( \frac{a_t^2}{a'_t^2} - \frac{a_t^2}{a'_t^2} \right)^2
\end{align*}
\]

Therefore,
\[
\frac{1}{n} \sum_{t=1}^{n} \left[ \frac{1}{2} \log \left( \frac{a_t^2}{a'_t^2} \right) \right] + \frac{1}{2} \frac{a_t^2}{a'_t^2} - \frac{a_t^2}{a'_t^2} = \frac{1}{4} \left[ \left( \frac{a_0^2}{a'_0^2} \right)^2 + \frac{1}{2} \log \left( \frac{a_0^2}{a'_0^2} \right) \right] + \frac{1}{2} \left( \frac{a_t^2}{a'_t^2} - \frac{a_t^2}{a'_t^2} \right)^2
\]

(A.3)

since \( a_t^2 = \frac{1}{n} \sum_{t=1}^{n} \frac{a_t^2}{a'_t^2} / \sqrt{n} \). As a result,
\[
\frac{1}{n} \sum_{t=1}^{n} \left[ \frac{1}{2} \log \left( \frac{a_t^2}{a'_t^2} \right) \right] + \frac{1}{2} \frac{a_t^2}{a'_t^2} - \frac{a_t^2}{a'_t^2} = \frac{1}{4} \left[ \left( \frac{a_0^2}{a'_0^2} \right)^2 + \frac{1}{2} \log \left( \frac{a_0^2}{a'_0^2} \right) \right] + \frac{1}{2} \left( \frac{a_t^2}{a'_t^2} - \frac{a_t^2}{a'_t^2} \right)^2
\]

(A.4)

The proof is complete once we notice that \( T_{rg} = n^{-1/2} \ln(\hat{\Theta}, \hat{\Theta}_g) / \sqrt{n} \) and apply Vuong (1985, Theorem 5.2).

**Proof of Lemma 4.2:** (i) follows from Vuong (1985, Lemma 4.1) by noticing that the conditions of that lemma are satisfied under A2 - A5.

To prove (ii), note from (4.6) that \( H^w_0 \) is equivalent to:
\[
[x'(a^* - y^*) + z' \beta^* - w' \delta^*] [2y - x'(a^* + y^*) - z' \beta^* - w' \delta^*] = 0
\]

\( H^w \)-almost surely. But given \( A_2 \),
\[
\text{Pr}\{ (y,z) : 2y - x'(a^* + y^*) + z' \beta^* + w' \delta^* = 0 \} = 0.
\]

Thus \( \text{Pr}\{ (y,z) : x'(a^* - y^*) + z' \beta^* - w' \delta^* = 0 \} = 1 \) which implies by A3 that \( \beta^* = \delta^* = 0 \) and \( a^* = y^* \). To prove that (ii) implies \( H^w_0 \), we note that when \( \beta^* = 0 \):
\[
\begin{bmatrix}
\sum_{xx} \\
\sum_{xz}
\end{bmatrix}
\begin{bmatrix}
a^* \\
0
\end{bmatrix} = \\
\begin{bmatrix}
\sum_{xy} \\
\sum_{zy}
\end{bmatrix}
\]

(see Lemma 2.1). Hence \( a^* = (\sum_{xx})^{-1} \sum_{xy} \). Similarly, we can show that \( y^* = (\sum_{xx})^{-1} \sum_{xy} \) when \( \delta^* = 0 \). Therefore \( e_f = y - x'a^* = y - x'y^* = e_g \). \( H^w \)-almost surely.

To prove (iii), we show that (iii) and (ii) are equivalent. If \( \beta^* = \delta^* = 0 \), then \( a^* = (\sum_{xx})^{-1} \sum_{xy} = y^* \) and \( \sum_{xx} a^* = \sum_{xy} \) from (A.5).

Similarly, \( \sum_{wx} \lambda^* = \sum_{wy} \). It can be easily shown that
\[
0, x^*, z^*, \lambda^w = (a^*, 0, 0) \text{ is the unique solution for}
\]
\[
\begin{bmatrix}
\sum_{xx} \\
\sum_{xz}
\end{bmatrix}
\begin{bmatrix}
\lambda_x \\
\lambda_z
\end{bmatrix} = \\
\begin{bmatrix}
\sum_{xy} \\
\sum_{zy}
\end{bmatrix}
\]

(A.6)

Hence \( (\lambda_x^*, \lambda_z^*, \lambda_w^*) = (a^*, 0, 0) \) from Lemma 2.1. Conversely, if \( \lambda_z^* = \lambda_w^* = 0 \), then we can easily show that \( (a, \beta) = (\lambda_x^*, 0) \) and \( (\gamma, \delta) = \)
(λ\*\*,0) are the unique solution for

\[ \begin{bmatrix} \sum_{xx} \sum_{xw} \\ \sum_{xx} \sum_{ZW} \end{bmatrix} \begin{bmatrix} a \\ \beta \end{bmatrix} = \begin{bmatrix} \sum_{xy} \\ \sum_{zy} \end{bmatrix} \]  

(A.7)

and

\[ \begin{bmatrix} \sum_{xx} \sum_{xw} \\ \sum_{wx} \sum_{WW} \end{bmatrix} \begin{bmatrix} \gamma \\ \delta \end{bmatrix} = \begin{bmatrix} \sum_{xy} \\ \sum_{zY} \end{bmatrix} \]  

(A.8)

respectively. Hence \( \beta^* = \delta^* = 0 \).

An example where \( \epsilon_\beta^2 = \epsilon_\delta^2 \neq \beta^* = \delta^* = 0 \): Consider the following two normal linear regression models for \( y \) given \( x = (z',w')' \) where \( z \) and \( w \) are both scalars:

\[ M_r: y = z\beta + \epsilon_r, \quad \epsilon_r \sim N(0,\sigma^2_r), \]

\[ M_g: y = w\delta + \epsilon_g, \quad \epsilon_g \sim N(0,\sigma^2_g). \]

Assume that the true joint p.d.f. of \((y,z,w)\) is:

<table>
<thead>
<tr>
<th>p.d.f</th>
<th>( w = -1 )</th>
<th>( w = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y = -6 )</td>
<td>( 2/12 )</td>
<td>( 0 )</td>
</tr>
<tr>
<td>( y = 0 )</td>
<td>( 0 )</td>
<td>( 1/4 )</td>
</tr>
<tr>
<td>( y = 6 )</td>
<td>( 1/12 )</td>
<td>( 0 )</td>
</tr>
<tr>
<td>( z = -1 )</td>
<td>( 0 )</td>
<td>( 1/4 )</td>
</tr>
<tr>
<td>( z = 1 )</td>
<td>( 1/4 )</td>
<td>( 0 )</td>
</tr>
</tbody>
</table>

That is, \( \Pr(y = -6,z = -1,w = -1) = 2/12 \) and so forth. Then

\[ \text{Var}_0(y,z,w) = \begin{bmatrix} 18 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \]

Hence, from Lemma 2.1, \( \beta^* = \delta^* = 1 \). But \( \Pr(\epsilon_\beta^2 = \epsilon_\delta^2) = \Pr(w - z)(2y - w - z) = 0 \) = 1. In fact, it can easily be shown that A3-A4 are satisfied; the only violation is A2.

**Proof of Lemma 4.3**: Given (4.19) and (4.20), the first and second partial derivatives of \( \log d_s(y|x_s;\theta_s) \) at \( \theta_s = \theta^*_s \) under \( H^0 \) are:

\[ \frac{\partial \log d_s(y|x_s;\theta^*_s)}{\partial \lambda_s} = \frac{x_s \epsilon}{\sigma^2} \]

\[ \frac{\partial \log d_s(y|x_s;\theta^*_s)}{\partial \sigma^2} = \frac{-2}{2\sigma^2} - \frac{1}{2\sigma^4} \]

\[ \frac{\partial^2 \log d_s(y|x_s;\theta^*_s)}{\partial \lambda^2_s} = \frac{x_s \epsilon^2}{\sigma^4} \]

\[ \frac{\partial^2 \log d_s(y|x_s;\theta^*_s)}{\partial \sigma^4} = \frac{-2}{2\sigma^4} + \frac{1}{2\sigma^6} \]

The lemma is then immediate upon replacing the above results into (4.16) – (4.18) and noting that \( E^0(\epsilon^2) = \sigma^2 \) and \( E^0(\epsilon x_s) = 0 \).

**Proof of Lemma 4.4**: Using the definition of \( W \) (i.e., equation (4.15)),
\[
\begin{vmatrix}
-\frac{1}{\sigma^2} C' \nu - \lambda I & \frac{1}{\sigma^2} U \\
\frac{1}{\sigma^2} U' \nu & \frac{1}{\sigma^2} C' \nu - \lambda I
\end{vmatrix} = \lambda^2 \begin{vmatrix}
-\frac{1}{\sigma^2} C' \nu & \frac{1}{\sigma^2} C' \nu - \lambda I \\
\frac{1}{\sigma^2} C' \nu & \frac{1}{\sigma^2} C' \nu - \lambda I
\end{vmatrix} = 0.
\]

\[\lambda^2 = \begin{vmatrix}
C_f - \lambda \sigma^2 \sum^0_{x_f x_f} C_{fg} \\
C_{fg} & C + \lambda \sigma^2 \sum^0_{x_f x_f} C_{fg}
\end{vmatrix} = 0. \quad (A.10)
\]

Upon substituting \(X_f = (x', z')\), \(X_g = (x', w')\) and applying the following (block) row and (block) column operations to (A.10):

(ii) subtract column 1 from column 3,

(iii) interchange column 3 and column 4, row 3 and row 4 respectively,

(iv) subtract row 1 from row 4,

(v) factorize \((\lambda \sigma^2)^k\) from column 4.

then \((A.10)\) implies

\[
\begin{vmatrix}
\sum^0_{xx} & \sum^0_{xz} & 0 \\
\sum^0_{xz} & \sum^0_{zz} & 0 \\
0 & 0 & -\sum^0_{wx}
\end{vmatrix} = 0.
\]

Postmultiplying by the non-singular matrix
\[
B = \begin{bmatrix}
I_k & -I_{xx} & I_{xz} & -I_{xx} & -I_{xx} & 0 \\
0 & I_p & 0 & 0 \\
0 & 0 & I_q & 0 \\
0 & 0 & 0 & I_k
\end{bmatrix}
\]

and premultiplying by \( B' \), we obtain:

\[
B'AB = \begin{bmatrix}
C_{xx} - \lambda \sigma^2 \sum_{xx} & C_{xz}R + \lambda \sigma^2 (0, \sum_{xz}) & \sum_{xx} \\
R'C_{xx} - \lambda \sigma^2 \sum_{xx} & R'CR - \lambda \sigma^2 \text{Diag}(Q_{zz}, -Q_{ww}) & 0 \\
\sum_{xx} & 0 & 0
\end{bmatrix}
\]

where \( R \) is given by (4.24) and \( C_{xx} = C'_{xx} \) is the first row block of \( C \).

Interchanging the first column block and the third column block, the matrix \( B'AB \) becomes block upper-triangular. Since \( \sum_{xx} \) is non-singular, then \( \det B'AB = 0 \) is equivalent to

\[
\lambda^2 \sigma^2 \cdot \det[R'CR - \lambda \sigma^2 \text{Diag}(Q_{zz}, -Q_{ww})] = 0.
\]

The proof is now complete.

**Proof of Proposition 4.5**: Immediate from Vuong (1985, Theorem 4.3).

**Proof of Lemma 4.6**: Part (i) follows from (3.3), (4.24), (4.28), and the partitioned inverse formula applied successively to \( \hat{\lambda}_f = (\hat{\alpha}, \hat{\beta})' \) and \( \hat{\beta}_g = (\hat{\gamma}, \hat{\delta})' \).

To prove (ii), we note that under \( H_0^b \):

\[
\frac{1}{\sqrt{n}} \sum_{t=1}^{n} X_t e_t \rightarrow N(0, C)
\]

by the multivariate central limit theorem, where \( C \) is given by (4.25) and \( e_t = \gamma_t - \frac{x_t a}{\gamma} = \gamma_t - \frac{x_t a}{\gamma} \). Moreover, we have:

\[
\sum_{xy} = \sum_{ax} + \frac{1}{n} \sum_{t=1}^{n} X_t e_t.
\]

Thus (4.34) becomes

\[
(\hat{\beta}', \hat{\alpha}') = \text{Diag}(Q_{zz}, Q_{ww}) \hat{\beta} + \frac{1}{n} \sum_{t=1}^{n} X_t e_t
\]

where the second equation follows from \( \hat{\beta}'_{XX} = 0 \). Part (ii) follows by applying (A.12).
Proof of Proposition 4.7: Immediate from Vuong (1986, Theorems 1 and 2).

Proof of Proposition 4.8: Part (i) follows from (3.3), (4.24), (4.28), and the partitioned inverse formula applied to 
\[ \hat{A} = (\hat{A}^t, \hat{A}_z, \hat{A}_w)'. \]
Since (A.13) holds under H\(\nu\)'s, then (4.38) becomes:
\[ (\hat{\lambda}_z, \hat{\lambda}_w)' = Q^{-1} \hat{R} \sum_{x} \hat{N} \hat{R} \hat{N}^{-1} = 0. \] Part (ii) follows by applying (A.12).

Proof of Proposition 4.9: Because \(\text{Diag}(Q_{zz}, Q_{ww})\) is non-singular, it can easily be shown that a matrix \(G\) is a g-inverse of \(V\) if and only if it is of the form
\[ G = \text{Diag}(Q_{zz}, Q_{ww})[R'CR]^{-1}\text{Diag}(Q_{zz}, Q_{ww}), \]where \([R'CR]^{-1}\) is a g-inverse of \(R'CR\). Similarly, since \(Q\) is non-singular, a matrix \(H\) is a g-inverse of \(W\) if and only if it is of the form
\[ H = Q[R'CR]^{-1}. \]

Proof of Lemma 5.1: Given \(A\) and Corollary 2.2, we have \(\lambda^0 = \lambda^*\) and \(\sigma^2 = \sigma_{pV}^2\). The result follows from Lemma 4.2, (4.12) and (4.13).

Proof of Lemma 5.2: Upon substituting (5.3) into (4.30), we have
\[ \det \begin{bmatrix} (1 - \lambda)Q_{zz} & Q_{zw} \\ Q_{zw} & (1 + \lambda)Q_{ww} \end{bmatrix} = 0. \]
For any \(\lambda \neq 1\), the above equation reduces to
\[ |(1 - \lambda)Q_{zz}|(1 - \lambda)Q_{zw}^{-1}Q_{zw}^{-1}Q_{ww} - (1 + \lambda)Q_{ww}| = 0, \]
i.e., \(|Q_{ww}^{-1}Q_{ww} - (1 - \lambda^2)Q_{ww}^{-1}Q_{ww}Q_{ww}^{-1}Q_{ww} - (1 + \lambda)Q_{ww}| = 0, \) which is (5.4). But (A.19) has p + q solutions, while (5.4) only has 2q solutions, hence the other p - q eigenvalues must be one. Moreover, from (5.4), the number of solutions \(\lambda^2 = 1\) which satisfy it is the same as the number of zero eigenvalues for \(Q_{ww}^{-1/2}Q_{ww}^{-1/2}Q_{ww}^{-1/2}\), which is in turn equal to rank \(Q_{ww}'\). Hence totally we have \(p + q - \text{rank} Q_{ww} = p - \text{rank} Q_{ww}\) one eigenvalues and \(q - \text{rank} Q_{ww}\) minus one eigenvalues. We now only have...
to show that $0 < \lambda^2 \leq 1$. First, since $(1 - \lambda^2)$ is the eigenvalue of a p.s.d. matrix $Q_{ww}^{-1/2} Q_{wz}^{-1} Q_{zw}^{-1/2}$, hence $\lambda^2 \leq 1$. Furthermore,

$$0 \neq \text{det} \begin{bmatrix} \sum_o x_{xx} & \sum_o x_{xz} & \sum_o x_{zx} \\ \sum_o z_{xx} & \sum_o z_{xz} & \sum_o z_{zx} \\ \sum_o w_{xx} & \sum_o w_{xz} & \sum_o w_{zx} \end{bmatrix} = \left| \sum_o x_{xx} \right| \text{det} \begin{bmatrix} Q_{zz} & Q_{zw} \\ Q_{wz} & Q_{ww} \end{bmatrix} = \left| \sum_o x_{xx} \right| |Q_{zz}| - Q_{wz}^{-1} Q_{zw}^{-1}$$

Therefore $\lambda_w = 0$ is not a solution for (5.4). The proof is now complete by using Lemma 4.4.

**Proof of Proposition 5.3:** Immediate from Proposition 4.5, Lemma 5.2, and the definition of a weighted sum of chi-squares.

**Proof of Corollary 5.4:** Obvious.

**Proof of Proposition 5.5:** The numerical equivalences between $W^1_n$ and $W^2_n$ follows from (A.18) since g-inverses need not be used. Moreover,

$$X_n = (\hat{R}^\prime C \hat{R})^{-1} = \sigma^2 (\hat{R}^\prime \hat{R})^{-1}.$$

But $\hat{R}^\prime \hat{R} = \hat{Q}$. This establishes (i).

Parts (ii) and (iii) follow from Propositions 4.7 and 4.9.

**Proof of Proposition 6.1:** Immediate from Lemma 5.2 and Vuong (1985, Theorem 6.4).

**Proof of Corollary 6.2:** Obvious.
may be violated much complicates the analysis.

7. The idea of discriminating between $M_f$ and $M_g$ by testing $e^*_g = e^*_f$ dates back to Hotelling (1940), who proposed, for the single explanatory variable case, a test of the more restrictive hypothesis that the correlation coefficients $\rho = \rho$ under $Y_{X_f}$, $Y_{X_g}$

the additional assumption that the true conditional distribution of $y$ given $(x_{x_f}, x_{x_g})$ is normal with linear conditional mean and constant conditional variance. For a generalization to more than one explanatory variables, see Chow (1980).

8. It can be shown that $\Delta^2$, $\Delta c^*$, and $\Delta F$ are biased estimators of $\text{AMSE}(M_f, M_g) = e^*_g - e^*_f$ even when $M_f$ and $M_g$ are both correctly specified. On the other hand, it can be shown that the statistic \[ \frac{\sigma^2_e + \sigma^2_{e_f}(n - l)}{\sigma^2_e + \sigma^2_{e_f}(n - l)} \] and the statistic \[ \frac{\sigma^2_e}{n - l} \] are both unbiased estimator of $e^*_g - e^*_f$ for fixed explanatory variables, the former under the assumption that $M_f \cap M_g$ is correctly specified, and the latter under the assumption that $M_f$ and $M_g$ are correctly specified.

9. For a generalization of Sawa criterion to non-linear models, see Chow (1981).

10. As the proof shows, the importance of Assumption A2 is to ensure that $e^2_f = e^2_g$ $p = 0$-almost surely is equivalent to $e^2_f = e^2_g$ $p = 0$-almost surely.

11. To ensure that these matrices exist, we assume, in addition to A3, that all fourth moments of the vector $(y, x')'$ exist.

12. The matrix $W$ simplifies if the competing models are asymptotically orthogonal, i.e., if $B_{f g} = 0$. Unfortunately, it can be shown that normal linear regression models can never be asymptotically orthogonal.

13. The exact number of zero eigenvalues of $W$ is $2k + 2 + (p + q - \text{rank} H^\prime CR)$. The non-zero eigenvalues of $W$ are the non-zero eigenvalues of $(\sigma^*_e)^{-2} H^\prime CR \text{Diag}(Q^{-1}_{zz}, Q^{-1}_{ww})$.

14. This ensures that, for any choice of $g$-inverse $G$ and any consistent sequence $[G_n]$, the statistic $W_n^1$ is asymptotically chi-square distributed under the null hypothesis $\beta^* = \delta^* = 0$ (see Vuong (1986)). An alternative and more frequent method is to use $(V_n^*)$ in place of $G_n$ where $V_n$ is a consistent estimator of $V$. Additional conditions on $V_n$ must, however, be imposed to insure the limiting chi-square distribution (see Andrews (1986) for the difficulties associated with this latter method.)

15. From Vuong (1986), the Wald statistics (4.36) based on different choice of $g$-inverses $G$ are asymptotically equivalent under $H^u_0$.

16. As the proof of Proposition 4.9 shows, for some particular choices of $G, H, G_n, H_n, W_n^1, W_n^2$ are numerically equal.

17. While $Q_{zw} = 0$ greatly simplifies the variance test and the LR test for model selection discussed in this and the next sections, this condition is precisely the one under which the Cox type tests cannot be applied (see, e.g., Pesaran (1974, p.158), Pesaran and Deaton (1978, p. 681), Davidson and MacKinnon (1981,
p. 785), and White (1982, p. 318).

18. If the larger model is correctly specified, \( \beta^* = \beta^0 \) and we have the classical hypothesis testing under correct specification.

19. To test \( H_0^0 \) against \( H_A^0 \), Vuong (1985, Theorem 7.4) considers twice the LR statistic which is in general asymptotically distributed as a weighted sum of chi-squares. One can also consider White (1982a)'s robust Wald and LM statistics. Asymptotic comparison of the tests based on these statistics is left for future research.

20. For a survey see MacKinnon (1983). Other non-nested hypothesis tests are the J and P tests proposed by Davidson and MacKinnon (1981) which are in fact Cox-type tests using a simpler estimator of \( AKLIC(\theta, \gamma) \) than the one initially proposed by Cox (see White (1982b)).

REFERENCES


