INTERNATIONAL ECONOMIC REVIEW
Vol. 21, No. 1, February, 1980

USING LEAST SQUARES TO APPROXIMATE UNKNOWN REGRESSION FUNCTIONS*

BY HALBERT WHITE[1]

1. INTRODUCTION

In econometric practice, models of the form

$$Y_i = g(Z_i) + \varepsilon_i \qquad i = 1,\ldots, n \tag{1}$$

are frequently encountered, where Y_i is a dependent variable which one is interested in explaining, $g(Z_i)$ is an unknown function of the independent variable Z_i (which may in general be a vector), and ε_i is a random variable with $E(\varepsilon_i)=0$, $E(\varepsilon_i^2)=\sigma_\varepsilon^2$, and $E(\varepsilon_i Z_i)=0$. The properties of $g(Z_i)$ are often the main focus of interest, and the econometrician's problem is to determine these properties by some computationally convenient procedure.

A procedure common in the literature is to "approximate $g(Z_i)$ by a Taylor series expansion" of suitable degree (usually a first or second order approximation is chosen) and, "ignoring terms of higher order," estimate the parameters of the resulting polynomial. In a recent text, J. Cramer [1969, pp. 79–83] attempts to provide a mathematically rigorous justification for this procedure. The notion that ordinary least squares (OLS) provides a Taylor series approximation is particularly widespread in the literature concerning the estimation of production functions. For example, Denny and Fuss [1977, p. 406] make it the foundation of their recent article, "The Use of Approximation Analysis to Test for Separability and the Existence of Consistent Aggregates." This notion has also been adopted and used in Spady and Friedlander [1978], Mincer [1974, p. 90], Aghevli and Khan [1977] and Atkinson and Halvorsen [1976].

The purpose of the present study is to point out the *severe* limitations of the Taylor series approximation interpretation for OLS and to provide in its place an approximation interpretation with general validity. The Taylor series interpretation is appealing since it may be used to study the local properties (derivatives, elasticities) of a function; the results of Section 2 provide some very limited conditions (from the practicing econometrician's viewpoint) under which OLS may be used to this end. In Section 3 we provide general conditions under which OLS yields a well-defined *least squares approximation* to an unknown function. This approximation has optimal *prediction* properties. Although well-known to statisticians (for example, see H. Cramér [1946, pp. 302–304]), the nature of least

* Manuscript received July 22, 1977; revised September 4, 1978.

[1] The author has benefitted from the comments of Walter Oi, Lawrence Olson and two anonymous referees.

squares as an approximation is not sufficiently well understood by empirical economists. The properties of the least squares approximation lead to a natural test for specification error given in Section 4, based on weighted least squares (WLS). In Section 5, we consider the problem of heteroskedasticity and the appropriateness and effects of WLS in the presence of possible functional misspecification. Section 6 contains a summary and concluding remarks.

2. LEAST SQUARES AND THE TAYLOR APPROXIMATION

The inexactness of the Taylor approximation interpretation is evidenced by a lack of agreement about the point of expansion. Usually, the approximation is considered to be taken at the mean of the explanatory variables (Cramer [1969], Spady and Friedlander [1978]), but Denny and Fuss [1977] take the approximation at the value unity. The fact that the parameters estimated by OLS do not necessarily correspond to those of the Taylor series expansion at the mean may be easily seen by the example in Figure 1. $T(Z_i)$ is the first order Taylor series expansion of $g(Z_i)$ around $\bar{Z}$, the mean of the Z_i. $L(Z_i)$ is the ordinary least squares estimate obtained when Y_i is regressed on the Z_i and a constant. $L(Z_i)$ has a different slope and intercept and lies below $T(Z_i)$, as do most of the observations, Y_i. Nor do the OLS estimates necessarily coincide with the slope at unity. Although there may be some point at which the OLS estimates and the slope coincide, locating this point requires a knowledge of the unknown function.

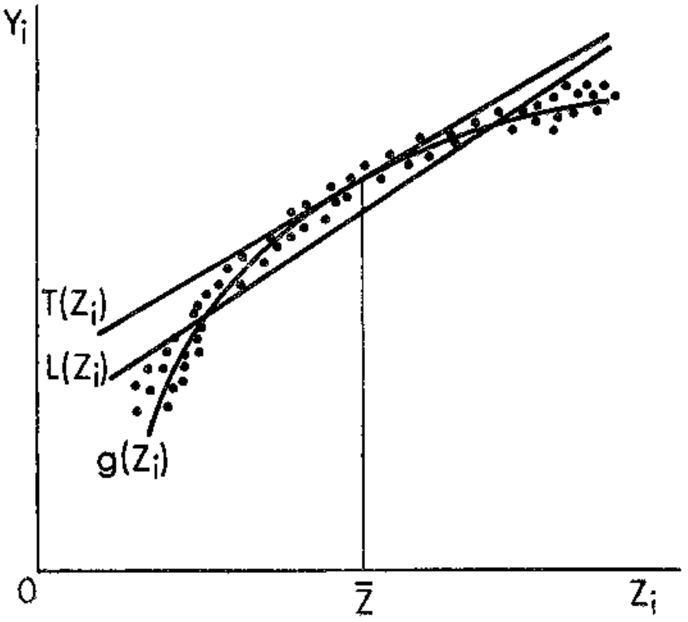

FIGURE 1

(In the many regressor case, or for higher order approximations, no such point need exist.)

An economic example will help to provide further insight. Suppose observations are generated by an unknown CES production function, say

$$\ln Q_i = -\ln(e^{-5\ln L_i} + 2e^{-5\ln K_i})/5 + \varepsilon_i \qquad i = 1, \ldots, n \tag{2}$$

where Q_i, L_i and K_i are output, labor and capital in period i (or for firm i) and ε_i is a random disturbance independent of L_i, K_i and distributed $N(0, \sigma_\varepsilon^2 = .01)$. Now consider the first and second order approximations in logarithms

$$\ln Q_i = \beta_0 + \beta_1 \ln L_i + \beta_2 \ln K_i \qquad i = 1, \ldots, n \tag{3}$$

$$\begin{aligned} \ln Q_i = \beta_0 + \beta_1 \ln L_i + \beta_2 \ln K_i + \beta_3(\ln L_i)^2 \\ + \beta_4 \ln L_i \cdot \ln K_i + \beta_5(\ln K_i)^2 \qquad i = 1, \ldots, n. \end{aligned} \tag{4}$$

Equation (3) is the familiar Cobb-Douglas production function and equation (4) is the increasingly popular translog production function introduced by Christensen, Jorgenson and Lau [1973].

Table 1 provides a numerical contrast between the results of ordinary least squares applied to (3) and (4) and the values of Taylor series coefficients of (2) evaluated at the means, $E(\ln L_i)$ and $E(\ln K_i)$. $\ln L_i$ and $\ln K_i$ are taken to be distributed uniformly on the unit interval. Thus $E(\ln L_i) = E(\ln K_i) = .5$. As a result of the properties of the true CES production function, the derivatives

TABLE 1

CONTRAST OF ORDINARY LEAST SQUARES PARAMETER ESTIMATES AND TAYLOR SERIES COEFFICIENTS

Dependent Variable: $\ln Q = -\ln(e^{-5\ln L} + 2e^{-5\ln K})/5 + \varepsilon$

	Cobb-Douglas Approximation	Translog Approximation	Taylor Approximation at the Means (.5)
const	−.3020 (.0233)†	−.2453 (.0345)	.2803
$\ln L$	.4286 (.0337)	.4606 (.1028)	.3333
$\ln K$	.5938 (.0359)	.6535 (.1201)	.6667
$(\ln L)^2$		−.4246 (.0993)	−2.2222
$\ln L \cdot \ln K$		.7922 (.0871)	−1.1111
$(\ln K)^2$		−.4800 (.1071)	−4.4444
s^2	.0184	.0122	
$\bar{R}^2$	.7320	.8225	

Number of observations: 200

† Specification robust standard errors in parentheses, computed from equation (7).

evaluated at any point where K and L are equal are the same — only the constant of the Taylor expansion is affected. Thus, the remarks which follow also apply to the expansion point chosen by Denny and Fuss [1977].

For the Cobb-Douglas specification, the coefficients of $\ln L$ and $\ln K$ are both at least two standard deviations away from the value of the Taylor series coefficients. The coefficient of $\ln K$ in the translog specification is fairly close to the corresponding Taylor coefficient, but that of $\ln L$ is algebraically even farther away than in the Cobb-Douglas case. The coefficients of the second order terms of the translog specification bear very little resemblance to the Taylor coefficients. The closest is eighteen standard deviations from the Taylor coefficient, and that of $\ln L \cdot \ln K$ has a statistically significant sign opposite that of the cross partial evaluated at the mean! These results indicate that tests of hypotheses based on Taylor approximation properties (as are those of Denny and Fuss [1977]) may be seriously misleading.

These examples illustrate the fact that OLS estimates do not necessarily provide reliable information about the local properties (derivatives, elasticities) of unknown functions. Results reported by Wales [1977] provide further related evidence, showing that estimated flexible functional forms such as the translog or generalized Leontief utility (production) functions do not necessarily provide good approximations to unknown functions in terms of their ability to satisfy the regularity conditions (quasi-concavity and monotonicity, which are derivative properties) required for utility maximization. Similar findings are reported by Griffin [1978].

If OLS is in general incapable of providing information about local properties of an unknown function, are there specific situations in which it can? In a recent article, Myers and Lahoda [1975] discuss optimal sample designs for using least squares to obtain information about the local properties of some particular unknown response functions. One requirement of these designs is that the regressors be orthogonal. This is cold comfort for the practicing econometrician who typically cannot design his sample and whose regressors are typically significantly correlated.[2]

Without the ability to control the experiment, the possibility of inferring properties of the unknown function still exists as Theorem 1 below shows. Unfortunately, the conditions of this result are very restrictive, but they are indicative of how far the Taylor series approximation interpretation may be taken. We make the following assumptions:

A1. The true model is

$$Y_i = g(Z_i) + \varepsilon_i \qquad i = 1, \ldots, n$$

where g is an unknown function and (Z_i, ε_i) are i.i.d. random $1 \times (p+1)$ vectors such that $E(Z_i)=0$, $E(Z_i' Z_i)=M_{ZZ}$ finite and nonsingular, $E(\varepsilon_i)=0$, $E(\varepsilon_i^2)=\sigma_\varepsilon^2<\infty$,

[2] With a sufficiently large sample one might be able to "design" the sample by discarding observations not sufficiently close to the design points, perhaps at the cost of some loss in efficiency.

$E(Z_i'\varepsilon_i)=0$, and $E(g(Z_i)^2)=\sigma_g^2<\infty$.

A2. g is twice differentiable such that: $\nabla g(0)$, the gradient evaluated at the origin, is bounded; the Hessian $\nabla^2 g(z)=\{\partial^2 g(z)/\partial z_i \partial z_j\}$ is positive semi-definite (p.s.d.) and there exist finite p.s.d. matrices A and B such that $A-\nabla^2 g(z)$ and $\nabla^2 g(z)-B$ are p.s.d. for all z in the support of F, the joint distribution function of Z_i.

Consider the linear approximation

$$Y_i = Z_i\theta + e_i \qquad i=1,\dots,n \tag{5}$$

where θ is a $p\times 1$ vector and define the least squares estimator $\hat\theta=(Z'Z)^{-1}Z'Y$ where Z is the $n\times p$ matrix with rows Z_i and Y_i is the $n\times 1$ vector with elements Y_i. Also, define the $p\times 1$ vectors γ_l and γ_u with elements

$$\gamma_{lj} = \frac{1}{2}\int_{-\infty}^{0}\int z_j zAz'dF(z) + \frac{1}{2}\int_{0}^{\infty}\int z_j zBz'dF(z)$$

$$\gamma_{uj} = \frac{1}{2}\int_{-\infty}^{0}\int z_j zBz'dF(z) + \frac{1}{2}\int_{0}^{\infty}\int z_j zAz'dF(z) \qquad j=1,\dots,p$$

where F is the joint distribution function of Z_i, A and B are as defined in A2, and by convention the first integral corresponds to the j-th element of z and the second represents the $p-1$ iterated integrals corresponding to the remaining elements of z. We adopt the notation $\geqq$ to indicate that a given matrix or vector equals or exceeds another, element by element. It is easily shown that $\gamma_u \geqq \gamma_l$ under A2. The following result provides conditions under which the OLS estimator can be used to provide information asymptotically about directional derivatives of the unknown function evaluated at the mean of the regressors.

THEOREM 1. *If A1 and A2 hold, and if $d'M_{ZZ}^{-1}\geqq 0$ where d is a $p\times 1$ direction vector of unit length, then*

$$d'\nabla g(0) + d'M_{ZZ}^{-1}\gamma_l \le \operatorname{plim} d'\hat\theta \le d'\nabla g(0) + d'M_{ZZ}^{-1}\gamma_u.$$

For example, suppose we are interested in bounding the j-th partial derivative evaluated at the mean, $g_j(0)$. The condition $d'M_{ZZ}^{-1}\geqq 0$ for this direction requires that the j-th row of M_{ZZ}^{-1} contain *only nonnegative elements*, a rather restrictive condition. When this condition fails, a bound need not exist. Note also that the bound becomes tighter, the closer to zero are the off-diagonal elements of the j-th row of M_{ZZ}^{-1}. The best possible situation for examining the gradient would be M_{ZZ}^{-1} diagonal, as in the Myers and Lahoda [1975] designs.

Another limitation of this result is that g is restricted by A2 to a subset of the concave functions. (A result similar to Theorem 1 is available for $\nabla^2 g$ negative semi-definite — a subset of the convex functions). This is perhaps not as serious a restriction as the requirement that $d'M_{ZZ}^{-1}\geqq 0$, since economic theory often justifies concavity assumptions. Also, by suitably restricting the range of the regressors, local concavity may be obtainable. Another limitation of this result is that one must have some idea of the matrices A and B. However, as we will see below, as the difference A–B becomes smaller, the bound becomes tighter, so it is not necessarily concavity *per se*, but differences in concavity which are important for the bound. To examine the implications of Theorem 1 more easily, consider the univariate case, $p=1$ (which also covers the case where Z_i has independent elements). In this case, the bound can be written

$$\begin{aligned} g'(0) + bE(Z_i^3)/2\sigma_Z^2 + (a-b)\delta_l \leq \operatorname{plim} \hat{\theta} \leq g'(0) \\ + bE(Z_i^3)/2\sigma_Z^2 + (a-b)\delta_u \end{aligned}$$

where a and b are the scalar analogs of A and B, $\sigma_Z^2 = E(Z_i^2)$ (recall $E(Z_i)=0$), and

$$\delta_l = (2\sigma_Z^2)^{-1} \int_{-\infty}^{0} z^3 dF(z)$$

$$\delta_u = (2\sigma_Z^2)^{-1} \int_{0}^{\infty} z^3 dF(z).$$

Now suppose the unknown function is quadratic in z, so that $a=b$. Then $\hat{\theta}$ provides an inconsistent estimate of $g'(0)$, unless Z_i has zero third moment. The extent of the inconsistency depends positively on the degree of concavity and the amount of skewness, and inversely on the variance of Z_i. In general, skewness will always introduce inconsistency. Now suppose $E(Z_i^3)=0$. Then $\delta_l = -\delta_u$, so that the bounds are symmetric, and the bounds becomes tighter as $a-b \to 0$. Without skewness, the bounds depend only on differences in concavity, not the amount of concavity. Finally, note that if we write δ_u more explicitly as

$$\delta_u(Z_i) = \frac{1}{2} \int_0^{\infty} z^3 dF(z) \Big/ \int_{-\infty}^{\infty} z^2 dF(z)$$

it can be shown that $\delta_u(\phi Z_i) = \phi\delta_u(Z_i)$ where $\phi > 0$ is a scale parameter. (Similarly, $\delta_l(\phi Z_i) = \phi\delta_l(Z_i)$). This implies that the precision of the bound improves as the range of the regressors decreases. This result is entirely analogous to a result considered "disturbing" by Wales [1977]: he found that flexible functional forms do better in mimicking the derivative properties of the unknown function when the range of the regressors is small. Specifically, as $\phi \to 0$, we have $\delta_u, \delta_l \to 0$ and $bE(Z_i^3)/2\sigma_Z^2 \to 0$; since $a-b$ cannot increase, $\operatorname{plim} \hat{\theta} \to g'(0)$, an entirely natural result. The efficiency considerations which led Wales to his reaction are only of secondary importance here.

Theorem 1 is a possibility theorem. Its conditions are sufficient for obtaining information about the gradient, but, when violated, there is still the chance that the least squares estimate will not be too far from the gradient. However, as the examples of this section demonstrate, this possibility is a slender straw on which to rest inferences about local properties of an unknown function.

3. THE LEAST SQUARES APPROXIMATION

In this section we will continue to assume A1, but A2 will be dropped. Now consider the linear approximation

$$Y_i = X_i\beta + u_i \qquad i = 1,\ldots, n \tag{5}$$

where $u_i \equiv g(Z_i) - X_i\beta + \varepsilon_i$ is a random variable which includes both the error of the approximation and the true stochastic error. The $1 \times k$ vector X_i has elements each of which is a function only of Z_i, but not necessarily a function of every element of Z_i — some variables may be omitted.

Suppose we wish to approximate (predict) Y_i for an observed vector X_i. If the occurrence of Y_i is not the result of a controlled experiment, but is determined by a random drawing from $F_{Z,\varepsilon}$, the joint distribution function of Z_i, ε_i, it is natural to consider minimizing the mean squared error (MSE) of approximation (prediction)

$$\sigma^2(\beta) = \int [g(z) - x\beta + \xi]^2 dF_{Z,\varepsilon}(z, \xi) \tag{6}$$

provided $\sigma^2(\beta)$ exists. To ensure this, we assume in addition to A1

A3. g and x are measurable functions of z.

Also we assume[3]

A4. $E(g(Z_i)\varepsilon_i) = 0$, $E(X_i'\varepsilon_i) = 0$, $E(X_i'X_i) = M_{XX}$, finite and nonsingular.

Define the OLS estimator $\hat{\beta}_{\text{OLS}} = (X'X)^{-1}X'Y$ where X is the $n \times k$ matrix with rows X_i. The next result provides conditions under which $\hat{\beta}_{\text{OLS}}$ converges asymptotically to β^*, the vector which minimizes the approximation (prediction) MSE.

THEOREM 2. *Under A1, A3 and A4, $\hat{\beta}_{\text{OLS}} \xrightarrow{\text{a.s.}} \beta^*$, the parameter vector which uniquely solves*

$$\min_{\beta} \sigma^2(\beta) = \int [g(z) - x\beta]^2 dF(z) + \sigma_\varepsilon^2$$

and $s^2 \xrightarrow{\text{a.s.}} \sigma^2(\beta^*)$ where $s^2 = (n-k)^{-1} \sum_{i=1}^{n} (Y_i - X_i\hat{\beta}_{\text{OLS}})^2$.

The vector β^* is the parameter vector of a *least squares approximation* $x\beta^*$ to the unknown function $g(z)$ with weighting function $dF(z)$. The properties of this approximation are well known to mathematicians.[4] If $g(z) \equiv x\beta_0$, then β^*

[3] The first two conditions of A4 result in no loss of generality. Since X_i need not depend on all Z_i, and since Z_i may contain unobservables, we can set $\varepsilon_i \equiv 0$.

[4] A well developed branch of mathematics known as Approximation Theory deals with general problems of this kind. The least squares approximation is well treated in Rice [1969, Chapters 2 and 12].

$=\beta_0$ for any distribution of the Z_i. If $g(z)\neq x\beta_0$, β^* will depend crucially on the distribution of Z_i. The weighting function ensures that frequently drawn X_t will yield small approximation errors at the cost of larger approximation errors for less frequently drawn X_i. Theorem 2 implies that $\hat{\beta}_{\text{OLS}}$ will share the properties of β^* asymptotically. In particular, $\hat{\beta}_{\text{OLS}}$ will depend upon F when $g(z)\neq x\beta_0$, a fact which provides the basis for a test of functional misspecification considered in the next section.

The estimated residual variance s^2 provides a consistent estimator of the minimum approximation MSE, $\sigma^2(\beta^*)$, which consists of two parts: the MSE due to the misspecification, $\int[g(z)+x\beta^*]^2dF(z)$, plus the error variance, σ_ε^2. s^2 can therefore be used to choose among competing models as advocated by Theil [1971, pp. 543–544] and Kloek [1975], since, when the regressors are orthogonal to ε_i, the smallest possible value for $\sigma^2(\beta^*)$, i.e. σ_ε^2, occurs when the model is correctly specified.[5]

The approximation result of Theorem 2 is not only more precise, but is much more general than that of Theorem 1: no restrictions are placed on M_{XX}^{-1} as they are on M_{ZZ}^{-1} by the requirement $d'M_{ZZ}^{-1}\geqq 0$; g is required only to be measurable rather than twice differentiable with p.s.d. Hessian everywhere; and the regressors X_i may be aggregates of underlying Z_i, may leave some Z_i entirely out of account, or may measure some Z_i subject to error, compared to Theorem 1, which requires the regressors to be precisely the Z_i. Generally, we can obtain good *predictions* of the level of Y_i using OLS; to get information about the gradient appears to require either the ability to design one's sample, or the severe restrictions on g and the distribution of Z_i imposed by Theorem 1.

With a large enough sample and an additional condition, $\hat{\beta}_{\text{OLS}}$ is approximately normally distributed, as the next result shows.

THEOMEM 3. *Under A1, A3 and A4*

$$\sqrt{n}(\hat{\beta}_{\text{OLS}}-\beta^*)\xrightarrow{\mathcal{L}} N(0,\ M_{XX}^{-1}V(\beta^*)M_{XX}^{-1})$$

provided $E(Y_i^2X_i'X_i)$ and $E(X_{ij}^2X_i'X_i)$, $j=1,\ldots,k$ are finite. Moreover, $(X'X/n)$ $\xrightarrow{\text{a.s.}} M_{XX}^{-1}$ and $\hat{V}_{\text{OLS}}=n^{-1}\sum_{i=1}^{n}(Y_i-X_i\hat{\beta}_{\text{OLS}})^2X_i'X_i\xrightarrow{\text{a.s.}}V(\beta^*)$, so that

$$(X'X/n)^{-1}\hat{V}_{\text{OLS}}(X'X/n)^{-1}\xrightarrow{\text{a.s.}}M_{XX}^{-1}V(\beta^*)M_{XX}^{-1}. \tag{7}$$

Note that the covariance matrix estimator $(X'X/n)^{-1}\hat{V}_{\text{OLS}}(X'X/n)^{-1}$ differs from the usual form $s^2(X'X/n)^{-1}$. This difference arises from the fact that with a misspecified model, $u_i^*\equiv g(Z_i)-X_i\beta^*+\varepsilon_i$ is only uncorrelated with X_i, not independent, so that V cannot be further factored. $s^2(X'X/n)^{-1}$ will be consistent if no misspecification has occurred and if X_i is independent of ε_i, in which case $\hat{V}_{\text{OLS}}\xrightarrow{\text{a.s.}}\sigma_\varepsilon^2M_{XX}$. Thus, a large difference between $\hat{V}_{\text{OLS}}$ and $s^2(X'X/n)$ can be

[5] See White and Olson [1979] for a formal test of the hypothesis that two alternative models with different regressors have the same prediction MSE.

evidence for misspecification.[6]

4. A TEST FOR FUNCTIONAL MISSPECIFICATION

As observed previously, the least squares approximation vector β^* generally depends on the weighting function $dF(z)$ when $g(z) \not\equiv x\beta_0$. When $g(z) \equiv x\beta_0$, $\beta^* = \beta_0$ regardless of the weighting function. The test for functional misspecification given in this section exploits these facts. The procedure involves first obtaining the usual estimator $\hat{\beta}_{\text{OLS}}$. Next, choose an arbitrary weighting function $W(Z_i)$ and choose β to minimize

$$n^{-1} \sum_{i=1}^{n} (Y_i - X_i\beta)^2 W(Z_i).$$

The solution to this problem is the weighted least squares (WLS) estimator

$$\hat{\beta}_{\text{WLS}} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y$$

where Ω^{-1} is a diagonal matrix with diagonal elements $W(Z_i)$. If no misspecification has occurred, $\hat{\beta}_{\text{OLS}}$ and $\hat{\beta}_{\text{WLS}}$ should be about the same; the presence of misspecification is indicated by $\hat{\beta}_{\text{OLS}}$ and $\hat{\beta}_{\text{WLS}}$ too far apart. Thus a test may be based on the differences $\hat{\beta}_{\text{OLS}} - \hat{\beta}_{\text{WLS}}$.

To obtain a test statistic we assume

A5. The weighting function W is a measurable function of z, $0 < \delta < W(z) < M$ for all z where δ, M are arbitrary finite constants, and $E(W(Z_i)X_i'\varepsilon_i) = 0$.

THEOREM 4. *If $g(z) \equiv x\beta_0$, if A1, A3–A5 hold, and if $E(Y_i^2 X_i'X_i)$, $E(X_{ij}^2 X_i'X_i)$, $j = 1, \ldots, k$ are finite, then*

$$n(\hat{\beta}_{\text{OLS}} - \hat{\beta}_{\text{WLS}})'\hat{\psi}^{-1}(\hat{\beta}_{\text{OLS}} - \hat{\beta}_{\text{WLS}}) \xrightarrow{A} \chi_k^2 \tag{8}$$

where

$$\hat{\psi} = (X'X/n)^{-1}\hat{V}_{\text{OLS}}(X'X/n)^{-1} + (X'\Omega^{-1}X/n)^{-1}\hat{V}_{\text{WLS}}(X'\Omega^{-1}X/n)^{-1} - (X'X/n)^{-1}\hat{U}(X'\Omega^{-1}X/n)^{-1} - (X'\Omega^{-1}X/n)^{-1}\hat{U}(X'X/n)^{-1}, \tag{9}$$

$$\hat{V}_{\text{WLS}} = n^{-1} \sum_{i=1}^{n} W(Z_i)^2 (Y_i - X_i\hat{\beta}_{\text{WLS}})^2 X_i'X_i,$$

$$\hat{U} = n^{-1} \sum_{i=1}^{n} W(Z_i)(Y_i - X_i\hat{\beta}_{\text{OLS}})(Y_i - X_i\hat{\beta}_{\text{WLS}})X_i'X_i,$$

provided a.s. $\lim \hat{\psi}$ *is nonsingular.*

If the statistic (8) is larger than the critical value for a χ_k^2 statistic at the α level, then the null hypothesis of no misspecification, $g(z) \equiv x\beta_0$, must be rejected at the α level (provided of course that the remaining conditions hold). Note that

[6] See White [1979] for a formal test of the hypothesis that $\sigma_*^2 M_{XX} = n^{-1} \sum_{i=1}^{n} E(u_i^{*2} X_i'X_i)$. If rejected, misspecification is implied.

the covariance matrix $\hat{\psi}$ given by (9) is correct under the alternative that $g(z) \not\equiv x\beta_0$. Under the null hypothesis, the estimator $\hat{\psi}$ is substantially simpler, as Hausman [1978] has shown. In this case, OLS is best asymptotically normal (BAN) while WLS is relatively inefficient asymptotically. Using Hausman's result, a consistent estimator is

$$\hat{\psi} = -s^2(X'X/n)^{-1} + s^2_{\mathrm{WLS}}(X'\Omega^{-1}X/n)^{-1}(X'\Omega^{-2}X/n)(X'\Omega^{-1}X/n)^{-1}$$

where $s^2_{\mathrm{WLS}} = (n-k)^{-1}\sum_{i=1}^{n} W(Z_i)(Y_i - X_i\hat{\beta}_{\mathrm{WLS}})^2$, provided ε_i is independent of X_i. The convenience of this form may outweigh possible small power losses from a practical viewpoint.

For fixed n, the power of this test will depend on the choice of $W(z)$. In such a situation it would seem appropriate to experiment with several different choices for $W(z)$. This suggestion appears to cause a problem: with different choices for $W(z)$, the test statistics (8) will not generally be independent, so that the exact size of the test performed becomes difficult to determine. There are two possible remedies for this difficulty. First, it is possible to stack the vectors $\hat{\beta}_{\mathrm{OLS}} - \hat{\beta}_{\mathrm{WLS}_1}, \ldots, \hat{\beta}_{\mathrm{OLS}} - \hat{\beta}_{\mathrm{WLS}_q}$ (corresponding to q different weighting schemes), obtain a giant covariance matrix which explicitly takes account of covariances between $\hat{\beta}_{\mathrm{OLS}} - \hat{\beta}_{\mathrm{WLS}_r}$ and $\hat{\beta}_{\mathrm{OLS}} - \hat{\beta}_{\mathrm{WLS}_s}$, $r, s = 1, \ldots, q$, and obtain a χ^2_{kq} statistic analogous to (8). The only thing to recommend this brute force technique is its precision. Alternatively, one may adopt the simpler procedure of accepting the null hypothesis if each of the q tests accepts the null hypothesis at the α level. Although the size of test resulting from this procedure is difficult to determine precisely, it is easily shown[7] to be bounded above by $q\alpha$. Thus, if any one of five tests rejects the null hypothesis at the 1% level, the null hypothesis may be rejected overall at or beyond the 5% level.

If $q\alpha$ is fixed, one still faces the problem of determining q (or α). Theoretically, the choice of q should be determined by power considerations: increasing q beyond a certain limit can lower the power of the test with fixed $q\alpha$, since the gain in power of having yet another check on the specification is paid for by an increase in the size ($q\alpha$) of the test (for fixed α). As a practical matter, the exact nature of this trade-off will be difficult to determine, but, considering the consistency of the test (power one for sufficiently large samples), it may be enough with large samples to try relatively few weighting schemes, say $q = 2$ or 3 in order to obtain sufficient power.

One possible weighting scheme is to use weighting functions of the form $(X_i\delta)^{-2}$, where δ is a fixed $1 \times k$ vector.[8] This amounts to weighting each observation by $(X_i\delta)^{-1}$ and then performing ordinary least squares. Prais and Houthakker [1971] perform such a weighted least squares estimation, although they chose $\delta = \hat{\beta}_{\mathrm{OLS}}$, not a fixed vector. Prais and Houthakker are particularly concerned with the problem of finding a good approximation to their demand functions, and it would be interesting to compare their OLS and WLS estimates using (8).

[7] See Lemma 1 of the Appendix.

[8] Provided, of course, that this choice satisfies A5.

Even though they provide a plausible theoretical justification for the heteroskedasticity which motivates their weighted least squares procedure, this heteroskedasticity was discovered by a thorough examination of the residuals. Less careful investigators may be tempted to embark on heteroskedasticity searches without a strong theoretical justification. What is discovered by regressing functions of the residuals on functions of the X_i may be evidence of functional misspecification rather than heteroskedasticity.[9] (See Section 5 below.) Further, when the functional form is misspecified and the ε_i are not truly heteroskedastic, there is no particular reason to prefer the WLS estimates. At least, the OLS estimates provide optimal predictions for a randomly drawn X_i. In situations where the investigator regresses functions of the residuals on various functions of the X_i, it may therefore be useful to compute the statistic (8).

As an example, (8) is computed for two different choices of $W(z)$ using the Cobb-Douglas and translog production function approximations considered in Section 2. For the first test, WLS_1, we choose $W(Z_i)=(X_i\delta)^{-2}$ where δ is taken to be the OLS estimator of Table 1. The second test, WLS_2, takes the weight to be .001 if $\ln L_i \leq .5$, unity otherwise. This is almost the same as dropping the observation if $\ln L_i \leq .5$. Table 2 compares the OLS and WLS_1 parameter estimates, and Table 3 compares the OLS and WLS_2 estimates.

In Table 2 we observe a moderate difference between $\hat{\beta}_{\text{OLS}}$ and $\hat{\beta}_{\text{WLS}_1}$ for both

TABLE 2

COMPARISON OF OLS AND WLS_1 PARAMETER ESTIMATES — $W_i=(X_i\hat{\beta}_{\text{OLS}})^{-2}$

Dependent Variable: $\ln Q$

	OLS Cobb-Douglas	WLS_1 Cobb-Douglas	OLS Translog	WLS_1 Translog
const	−.3020 (.0233)[†]	.0530 (.0997)	−.2453 (.0345)	−.0331 (.1196)
$\ln L$	.4284 (.0337)	−.0877 (.1880)	.4606 (.1028)	.2709 (.4283)
$\ln K$	.5938 (.0359)	.2992 (.2571)	.6535 (.1201)	.0878 (.3324)
$(\ln L)^2$			−.4246 (.0993)	−.6646 (.3488)
$\ln L \cdot \ln K$			.7922 (.0871)	1.9365 (.4656)
$(\ln K)^2$			−.4800 (.1071)	−.2641 (.2462)
Parameter difference test for misspecification	$\chi^2_3=13.1390$		$\chi^2_6=26.2877$	

[†] Specification robust standard errors from equation (7) in parentheses.

[9] This procedure is suggested by Glejser [1969].

TABLE 3
COMPARISON OF OLS AND WLS_2 PARAMETER ESTIMATES — $W_i=.001$ IF $\ln L_i \leq .5$, UNITY OTHERWISE
Dependent Variable: $\ln Q$

	OLS Cobb-Douglas	WLS_2 Cobb-Douglas	OLS Translog	WLS_2 Translog
const	−.3020 (.0233)†	−.3166 (.0749)	−.2453 (.0345)	−.6040 (.3967)
$\ln L$	.4284 (.0337)	.2918 (.0912)	.4606 (.1028)	1.0939 (1.0159)
$\ln K$	.5938 (.0359)	.8152 (.0389)	.6535 (.1201)	.9321 (.2582)
$(\ln L)^2$			−.4246 (.0993)	−.6214 (.6508)
$\ln L \cdot \ln K$			.7922 (.0871)	.2759 (.2634)
$(\ln K)^2$			−.4800 (.1071)	−.3392 (.1470)
Parameter difference test for misspecification	$\chi^2_3=39.9896$		$\chi^2_6=7.8084$	

† Specification robust standard errors from equation (7) in parentheses.

Cobb-Douglas and translog specifications. Appearances are confirmed by the χ^2 statistic (8). For each model, we must reject the null hypothesis that the model is correctly specified at beyond the $\alpha=.5\%$ level. In Table 3 we find WLS_2 estimates to be rather different from either OLS or WLS_1. Basing a formal χ^2 test on the difference between $\hat{\beta}_{\mathrm{OLS}}$ and $\hat{\beta}_{\mathrm{WLS}_2}$, we find that only for the Cobb-Douglas model can the null hypothesis of no misspecification be rejected at standard levels. If this test were used alone, a type II error would be committed for the translog specification.

Following the procedure suggested above of accepting the null hypothesis if both tests accept the null hypothesis at the α level leads to a test of size $\leq 2\alpha$. Thus, combining the results of Tables 2 and 3, we can reject the null hypothesis of no misspecification at or beyond the 1% level. Taken together, the tests perform well in detecting the functional misspecification.

For purposes of comparison, the covariance matrix of the OLS regression coefficients for the Cobb-Douglas specification is reported in Table 4, calculated first as $s^2(X'X)^{-1}$ and then using the specification robust estimator of Theorem 3. As previously noted, $s^2(X'X)^{-1}$ is generally inconsistent in the presence of misspecification. Here, differences between the usual and specification robust covariance estimators are on the order of 10%, with some elements increasing and some elements decreasing when $(X'X)^{-1}[\sum_{i=1}^{n}(Y_i-X_i\hat{\beta}_{\mathrm{OLS}})^2X_i'X_i](X'X)^{-1}$ is used instead of $s^2(X'X)^{-1}$. In White [1979], a formal test for model misspecification based on differences in these covariance matrix estimators is constructed. In this particular case, this test ($\chi^2_5=152.68$) reveals a statistically significant differ-

TABLE 4

CONTRAST OF COVARIANCE MATRICES OF OLS REGRESSIONS FOR THE COBB-DOUGLAS APPROXIMATION

1. Covariance matrix calculated as $s^2(X'X)^{-1}$

$$\begin{bmatrix} .000594 & -.000494 & -.000496 \\ -.000494 & .001049 & -.000076 \\ -.000496 & -.000076 & .001058 \end{bmatrix}$$

2. Covariance matrix calculated as $(X'X)^{-1}[\sum_{i=1}^{n} (Y_i - X_i\hat{\beta}_{\text{OLS}})^2 X_i'X_i] (X'X)^{-1}$

$$\begin{bmatrix} .000542 & -.000415 & -.000448 \\ -.000415 & .001132 & -.000389 \\ .000448 & -.000389 & .001292 \end{bmatrix}$$

ence between the two, so that use of $s^2(X'X)^{-1}$ for testing hypotheses will lead to errors in inference. (See White [1979] for further discussion and examples.)

5. HETEROSKEDASTIC ERRORS

When there is no functional misspecification and the errors ε_i are truly heteroskedastic, so that $E(\varepsilon\varepsilon') = \Omega \neq \sigma^2 I$, the WLS estimator

$$\hat{\beta}_{\text{WLS}} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y$$

is consistent and BAN. Above, we have seen that with misspecification the use of a weighting function may affect the parameter estimates obtained. In this section we show that correcting for a particular kind of heteroskedasticity[10] in the presence of functional misspecification does not impair one's ability to estimate the parameter vector of the least squares approximation, β^*.

Specifically, assume

A6. The weights W_i are i.i.d. random variables independent of Z_i, $0 < \delta < W_i < M$, where δ, M are arbitrary finite constants, and $E(W_i X_i' \varepsilon_i) = 0$.

Correcting for heteroskedasticity involves setting $W_i = 1/\sigma_i^2$, where $\sigma_i^2 = E(\varepsilon_i^2)$, $i = 1, \ldots, n$ are scalars independent of Z_i. Continuing to define Ω^{-1} as the diagonal matrix with diagonal elements W_i, we have the following result.

THEOREM 5. *If A1, A3, A4 and A6 hold, then* $\hat{\beta}_{\text{WLS}} \xrightarrow{\text{a.s.}} \beta^*$.

The independence of W_i and Z_i is crucial to the consistency of $\hat{\beta}_{\text{WLS}}$ for β^*.

In practice, the choice $W_i = 1/\sigma_i^2$ is unlikely to be known *a priori*. Usually, σ_i^2 must be estimated from the data. This may be successfully done if σ_i^2 depends only upon factors independent of Z_i. If this is not the case, the determinants of heteroskedasticity can be confounded with other terms arising from the non-independence of u_i^* and Z_i caused by functional misspecification. To see this, suppose (for simplicity), that β^* is known, but σ_i^2 is only known to satisfy $\sigma_i^2 = Q_i\lambda$

[10] The heteroskedasticity considered here involves ε_i with i.i.d. variances independent of Z_i.

where Q_i are i.i.d. random variables. If the model were correct, σ_i^2 could be estimated as $Q_i\hat{\lambda}$ where $\hat{\lambda}$ is the estimated regression coefficient of the model

$$\varepsilon_i^2 = Q_i\lambda + v_i \qquad i = 1,\ldots, n$$

and $\varepsilon_i = Y_i - X_i\beta_0 \equiv Y_i - X_i\beta^*$. When the model is misspecified, two things happen. First, the OLS regression for

$$u_i^{*2} = Q_i\lambda + v_i$$

where $u_i^* = Y_i - X_i\beta^* \equiv \varepsilon_i + g(Z_i) - X_i\beta^*$ will yield biased estimates of λ when Q_i is correlated with the squared specification error. Second, even if λ were known, dependence of Q_i on Z_i will affect the least squares approximation when $W_i = 1/Q_i\lambda$. Thus, attempting to correct for heteroskedasticity can eliminate even the modest properties of the least squares estimator as an optimal predictor (in the MSE sense) for randomly drawn X_i.

Can anything be done about this difficulty? Short of eliminating the misspecification, the answer appears to be no. However, this question presupposes that something *should* be done about heteroskedasticity. When the model is known, elimination of heteroskedasticity improves the efficiency of the parameter estimates. However, when the model is unknown, even a successful correction for heteroskedasticity (say $W_i = 1/\sigma_i^2$ known and independent of Z_i) need not necessarily reduce the covariance matrix, $M_{\tilde{X}\tilde{X}}^{-1}V(\beta^*)M_{\tilde{X}\tilde{X}}^{-1}$, because of the special form of $V(\beta^*)$. In the author's opinion, a safe strategy is to estimate β by OLS and WLS, correcting for suspected heteroskedasticity. Perform the specification test (8) based on a comparison of these two estimators. If the null hypothesis of no misspecification is rejected, use the OLS estimates for purposes of prediction (approximation). Note that the weights of A6 have no power for the specification test (8). If such weights are available, WLS may (but does not necessarily) improve the prediction (approximation) variance.

6. SUMMARY AND CONCLUDING REMARKS

Functional misspecification is a fact of life — one almost never has information which justifies a particular linear or non-linear specification. Indeed, most econometric estimating relationships are intended as approximations, rather than as the "truth." It is therefore useful to realize the limitations of our approximations. As predictors they have desirable properties. The parameters estimated converge to the parameters of the least squares approximation, however, and not to the "true" parameters except when no misspecification occurs. Inferences may be drawn about the parameters of the *approximation* when performing least squares in the presence of functional misspecification using the covariance estimator $(X'X/n)^{-1}\hat{V}_{\mathrm{OLS}}(X'X/n)^{-1}$. The usual estimator $s^2(X'X/n)^{-1}$ is not necessarily consistent in the framework considered here and may yield faulty inferences. When using weighted least squares it may not be possible to obtain consistent estimates of the parameters of the conditional error variances, due to the presence of the approximation remainder in the estimated ordinary least squares residuals. A test for specification error, (8), is provided which may be useful in assessing the extent of these difficulties.

It is worthwhile to emphasize the general usefulness of the least squares approximation. Its optimality as a predictor holds regardless of omitted variables, aggregation error, errors in variables, simultaneous equation error, non-additivity of the disturbance, or other forms of functional misspecification. This is in sharp contrast to the very limited ability of least squares to provide information about partial derivatives or elasticities of unknown functions, as the results of Section 2 suggest. Reliance on the Taylor approximation interpretation is an imprecise if not totally misleading practice.

University of Rochester, U.S.A.

MATHEMATICAL APPENDIX

All assumptions and definitions are as given in the text.

THEOREM 1. *If A1 and A2 hold, and if $d'M_{ZZ}^{-\frac{1}{2}} \geqq 0$, where d is a $p \times 1$ direction vector of unit length, then*

$$d'\nabla g(0) + d'M_{ZZ}^{-\frac{1}{2}}\gamma_l \leq \operatorname{plim} d'\hat{\theta} \leq d'\nabla g(0) + d'M_{ZZ}^{-\frac{1}{2}}\gamma_u.$$

PROOF. By the finiteness of M_{ZZ} and σ_g^2, the Hölder inequality ensures the finiteness of $E(Z_{ij}g(Z_i))$, so that we can write

$$E(Z_{ij}g(Z_i)) = \int_{-\infty}^{0}\!\!\int z_j g(z)dF(z) + \int_{0}^{\infty}\!\!\int z_j g(z)dF(z). \tag{a.1}$$

Using the mean value theorem and the boundedness of $\nabla g(0)$, we obtain

$$\int_{-\infty}^{0}\!\!\int z_j g(z)dF(z) = \int_{-\infty}^{0}\!\!\int z_j[g(0) + z\nabla g(0) + \frac{1}{2}\, z\nabla^2 g(\bar{z}(z))z']dF(z)$$

and

$$\int_{0}^{\infty}\!\!\int z_j g(z)dF(z) = \int_{0}^{\infty}\!\!\int z_j[g(0) + z\nabla g(0) + \frac{1}{2}\, z\nabla^2 g(\tilde{z}(z))z']dF(z)$$

where $\bar{z}(z)$ and $\tilde{z}(z)$ lie between 0 and z. By A2, $zBz' \leq z\nabla^2 g(\bar{z}(z))z' \leq zAz'$ and $zBz' \leq z\nabla^2 g(\tilde{z}(z))z' \leq zAz'$ for all z in the support of F. Hence

$$\begin{aligned} \int_{-\infty}^{0}\!\!\int z_j\left[g(0) + z\nabla g(0) + \frac{1}{2}\, zAz'\right]dF(z) &\leq \int_{-\infty}^{0}\!\!\int z_j g(z)dF(z) \\ &\leq \int_{-\infty}^{0}\!\!\int z_j\left[g(0) + z\nabla g(0) + \frac{1}{2}\, zBz'\right]dF(z) \end{aligned} \tag{a.2}$$

and

$$\text{(a.3)} \qquad \int_0^\infty\!\!\int z_j\left[g(0) + z\nabla g(0) + \frac{1}{2}\,zBz'\right]dF(z) \leq \int_0^\infty\!\!\int z_j g(z)dF(z)$$

$$\leq \int_0^\infty\!\!\int z_j\left[g(0) + z\nabla g(0)\frac{1}{2}\,zAz'\right]dF(z).$$

Combining (a.1), (a.2) and (a.3) and using the fact that $E(Z_{ij})=0$ yields

$$\text{(a.4)} \qquad \left[\int_{-\infty}^{\infty}\!\!\int z_j z dF(z)\right]\nabla g(0) + \gamma_{lj} \leq E(Z_{ij}g(Z_i))$$

$$\leq \left[\int_{-\infty}^{\infty}\!\!\int z_j z dF(z)\right]\nabla g(0) + \gamma_{uj}, \qquad j = 1,\dots, p.$$

Stacking the relations (a.4) yields

$$\text{(a.5)} \qquad M_{ZZ}\nabla g(0) + \gamma_l \leqq M_{Zg} \leqq M_{ZZ}\nabla g(0) + \gamma_u$$

where $M_{Zg}=E(Z_i'g(Z_i))$. Since $d'M_{ZZ}^{-1}\geqq 0$, the inequalities of (a.5) are not disturbed by pre-multiplication by $d'M_{ZZ}^{-1}$, so

$$d'\nabla g(0) + d'M_{ZZ}^{-1}\gamma_l \leq d'M_{ZZ}^{-1}M_{Zg} \leq d'\nabla g(0) + d'M_{ZZ}^{-1}\gamma_u.$$

The result follows if plim $d'\hat{\theta}=d'M_{ZZ}^{-1}M_{Zg}$. This is true since

$$\hat{\theta} = (Z'Z/n)^{-1}(Z'Y/n)$$

$$= (Z'Z/n)^{-1}(n^{-1}\sum_{i=1}^{n} Z_i'g(Z_i)) + (Z'Z/n)^{-1}(n^{-1}\sum_{i=1}^{n} Z_i'\varepsilon_i).$$

Now $(Z'Z/n)\xrightarrow{p} M_{ZZ}$ by Khintchine's weak law of large numbers and since M_{ZZ} is nonsingular, $(Z'Z/n)^{-1}\xrightarrow{p} M_{ZZ}^{-1}$. Also by Khintchine's weak law, $n^{-1}\sum_{i=1}^{n} Z_i'g(Z_i) \xrightarrow{p} M_{Zg}$ and $n^{-1}\sum_{i=1}^{n} Z_i'\varepsilon_i \xrightarrow{p} 0$ (since $E(Z_i'\varepsilon_i)=0$ under A1). Thus

$$\hat{\theta} \xrightarrow{p} M_{ZZ}^{-1}M_{Zg}^{-1}$$

and the result follows. □

Theorem 2. *Under A1, A3 and A4, $\hat{\beta}_{\mathrm{OLS}}\xrightarrow{a.s.}\beta^*$, the parameter vector which uniquely solves*

$$\min_{\beta} \sigma^2(\beta) = \int[g(z) - x\beta]^2 dF(z) + \sigma_\varepsilon^2$$

and $s^2\xrightarrow{a.s.}\sigma^2(\beta^*)$ where $s^2=(n-k)^{-1}\sum_{i=1}^{n}(Y_i-X_i\hat{\beta}_{\mathrm{OLS}})^2$.

Proof. Define $\beta^*=M_{XX}^{-1}M_{Xg}$ where $M_{Xg}=\int x'g(z)dF(z)$. The vector β^* is finite since M_{XX}^{-1} is finite and since M_{Xg} is finite by the Hölder inequality under A4. Consider a compact neighborhood of β^*, say ν. Now $\sigma^2(\beta)$ is integrable for $\beta^* \in \nu$, and there exist integrable functions $h_j(z)$, $j=1,\dots, k$ such that $|\partial[g(z)-x\beta]^2/$

$\partial\beta_j| = |x_j[g(z) - x\beta]| \leq h_j(z) = |x_j g(z)| + \sum_{l=1}^{k} |x_j x_l| \cdot b_l$ where b_l is a finite constant such that $|\beta_l| \leq b_l$ for all β in ν. It follows from Bartle [1966, Corollary 5.9] that $\sigma^2(\beta)$ is differentiable on ν and that

$$\partial\sigma^2(\beta)/\partial\beta = \int x'[g(z) - x\beta]dF(z).$$

It is readily verified that $\partial\sigma^2(\beta^*)/\partial\beta = 0$, and by a similar argument that $\partial^2\sigma^2(\beta)/\partial\beta\partial\beta' = M_{XX}$, which is positive definite by A4. Thus β^* uniquely minimizes $\sigma^2(\beta)$. Now

$$\begin{aligned}\hat{\beta}_{\text{OLS}} &= (X'X/n)^{-1}(X'Y/n)\\ &= (X'X/n)^{-1}(n^{-1}\sum_{i=1}^{n} X_i g(Z_i)) + (X'X/n)^{-1}(n^{-1}\sum_{i=1}^{n} X_i'\varepsilon_i)\end{aligned}$$

given A1. Since (Z_i, ε_i) are i.i.d., Komolgorov's strong law of large numbers implies $(X'X/n) \xrightarrow{\text{a.s.}} M_{XX}$, and the nonsingularity of M_{XX} guarantees $(X'X/n)^{-1} \xrightarrow{\text{a.s.}} M_{XX}^{-1}$. Also Komolgorov's strong law implies $n^{-1}\sum_{i=1}^{n} X_i g(Z_i) \xrightarrow{\text{a.s.}} M_{Xg}$, $n^{-1}\sum_{i=1}^{n} X_i'\varepsilon_i \xrightarrow{\text{a.s.}} 0$ under A3 and A4. It follows that

$$\hat{\beta}_{\text{OLS}} \xrightarrow{\text{a.s.}} M_{XX}^{-1}M_{Xg} = \beta^*.$$

Next,

$$\begin{aligned}\sigma^2(\beta^*) &= \int[g(z)^2 - 2\beta^{*\prime}x'g(z) + \beta^{*\prime}x'x\beta^*]dF(z) + \sigma_\varepsilon^2\\ &= \sigma_g^2 - 2M_{Xg}'M_{XX}^{-1}M_{Xg} + M_{Xg}'M_{XX}^{-1}M_{Xg} + \sigma_\varepsilon^2\\ &= \sigma_g^2 + \sigma_\varepsilon^2 - M_{Xg}'M_{XX}^{-1}M_{Xg}.\end{aligned}$$

Now $s^2 = (n-k)^{-1}\sum_{i=1}^{n}(Y_i - X_i\hat{\beta}_{\text{OLS}})^2$ may be written

$$s^2 = [n/(n-k)]\,[Y'Y/n - (Y'X/n)(X'X/n)^{-1}(X'Y/n)].$$

By Komolgorov's strong law of large numbers, $(Y'Y/n) \xrightarrow{\text{a.s.}} \sigma_g^2 + \sigma_\varepsilon^2$ given A1 and A4, and, as before, $(X'Y/n) \xrightarrow{\text{a.s.}} M_{Xg}$, $(X'X/n)^{-1} \xrightarrow{\text{a.s.}} M_{XX}^{-1}$. Since $n/(n-k) \to 1$,

$$s^2 \xrightarrow{\text{a.s.}} \sigma_g^2 + \sigma_\varepsilon^2 - M_{Xg}'M_{XX}^{-1}M_{Xg} = \sigma^2(\beta^*). \quad \square$$

THEOREM 3. *Under A1, A3 and A4*

$$\sqrt{n}(\hat{\beta}_{\text{OLS}} - \beta^*) \overset{A}{\sim} N(0, M_{XX}^{-1}V(\beta^*)M_{XX}^{-1})$$

where $V(\beta^*) = E([g(Z_i) - X_i\beta^* + \varepsilon_i]^2 X_i'X_i)$, *provided* $E(Y_i^2 X_i'X_i)$ *and* $E(X_{ij}^2 X_i'X_i)$ $j = 1, \ldots, k$ *are finite. Moreover,* $(X'X/n)^{-1} \xrightarrow{\text{a.s.}} M_{XX}^{-1}$ *and* $\hat{V}_{\text{OLS}} \xrightarrow{\text{a.s.}} V(\beta^*)$, *where* $\hat{V}_{\text{OLS}} = n^{-1}\sum_{i=1}^{n}(Y_i - X_i\hat{\beta}_{\text{OLS}})^2 X_i'X_i$ *so that*

$$(X'X/n)^{-1}\hat{V}_{\mathrm{OLS}}(X'X/n)^{-1} \xrightarrow{\text{a.s.}} M_{XX}^{-1}V(\beta^*)M_{XX}^{-1}.$$

PROOF. We may write

$$\sqrt{n}(\hat{\beta}_{\mathrm{OLS}} - \beta^*) = (X'X/n)^{-1}n^{-1/2}\sum_{i=1}^{n} X_i'(g(Z_i) - X_i\beta^* + \varepsilon_i).$$

The random vectors $X_i'(g(Z_i)-X_i\beta^*+\varepsilon_i)$ are i.i.d. under A1 with

$$E(X_i'(g(Z_i) - X_i\beta^* + \varepsilon_i)) = 0$$

and

$$E([g(Z_i) - X_i\beta^* + \varepsilon_i]^2X_i'X_i) = V(\beta^*)$$

given A1, A3, A4, provided $V(\beta^*)$ is finite. Thus,

$$n^{-1/2}\sum_{i=1}^{n} X_i'(g(Z_i) - X_i\beta^* + \varepsilon_i) \xrightarrow{A} N(0, V(\beta^*))$$

by the multivariate Lindeberg-Levy central limit theorem. Since the X_i are i.i.d., $(X'X/n)\xrightarrow{\text{a.s.}}M_{XX}$ by Komolgorov's strong law of large numbers under A4; the nonsingularity of M_{XX} thus implies $(X'X/n)^{-1}$ exists almost surely for n sufficiently large, and that $(X'X/n)^{-1}\xrightarrow{\text{a.s.}}M_{XX}^{-1}$. It follows that given A1, A3 and A4

$$\sqrt{n}(\hat{\beta}_{\mathrm{OLS}} - \beta^*) \xrightarrow{A} N(0, M_{XX}^{-1}V(\beta^*)M_{XX}^{-1})$$

provided $V(\beta^*)$ is finite. This fact and the strong consistency of $\hat{V}_{\mathrm{OLS}}$ for $V(\beta^*)$ are proven as follows.

By assumption, $E(Y_i^2X_i'X_i)$ and $E(X_{ij}^2X_i'X_i)$ are finite. Since β^* is finite, $E(X_{ij}^2\beta_j^{*2}X_i'X_i)$ is finite. Also, since $X_{ij}\beta_jX_i'X_i$ is continuous in β_j, there exists a compact neighborhood v of β^* such that $E(|X_{ij}^2\beta_j^2X_{il}X_{im}|)$ $j, l, m=1,\ldots, k$ is finite for all β in v. There also exists $\tilde{\beta}$ with finite elements $\tilde{\beta}_j$ such that $\beta_j^2\leq\tilde{\beta}_j^2$ for all β in v, so that

$$|X_{ij}^2\beta_j^2X_{il}X_{im}| \leq |X_{ij}^2\tilde{\beta}_j^2X_{il}X_{im}| \qquad j, l, m = 1,\ldots, k$$

and $|X_{ij}^2\tilde{\beta}_j^2X_{il}X_{im}|$ is integrable with respect to F (the joint distribution function of the Z_i). It is a direct consequence of the simple inequality $|a+b|^2\leq 2|a|^2+2|b|^2$ that there exist finite constants $\lambda_0,\ldots, \lambda_k$ such that for all β in v

$$\text{(a.6)} \qquad |(Y_i - X_i\beta)^2X_{il}X_{im}| \leq \lambda_0|Y_i^2X_{il}X_{im}| + \sum_{j=1}^{k}\lambda_j|X_{ij}^2\beta_j^2X_{il}X_{im}|$$

$$\leq \lambda_0|Y_i^2X_{il}X_{im}| + \sum_{j=1}^{k}\lambda_j|X_{ij}^2\tilde{\beta}_jX_{il}X_{im}| \qquad l, m = 1,\ldots, k.$$

Taking expectations at β^* in (a.6), the finiteness of $E(Y_i^2X_{il}X_{im})$ and $E(|X_{ij}^2\tilde{\beta}_j^2\cdot X_{il}X_{im}|)$ guarantees the finiteness of $V(\beta^*)$. The fact that $\hat{\beta}_{\mathrm{OLS}}\xrightarrow{\text{a.s.}}\beta^*$ (proven in Theorem (2) under A1, A3 and A4), the finiteness of $E(Y_i^2X_{il}X_{im})$ and $E(|X_{ij}^2\tilde{\beta}_j^2\cdot X_{il}X_{im}|)$ and (a.6) imply

$$n^{-1}\sum_{i=1}^{n}(Y_i - X_i\hat{\beta}_{\mathrm{OLS}})^2 X_{il}X_{im} \xrightarrow{\text{a.s.}} E([g(Z_i) - X_i\beta^* + \varepsilon_i]^2 X_{il}X_{im})$$

$$l,\ m = 1,\ldots,k$$

by White [1979, Lemma 3.1]; thus, $\hat{V}_{\mathrm{OLS}} \xrightarrow{\text{a.s.}} V(\beta^*)$. □

THEOREM 4. *If $g(Z_i) \equiv X_i\beta_0$, if A1, A3–A5 hold, and if $E(Y_i^2 X_i'X_i)$, $E(X_{ij}^2 \cdot X_i'X_i)$, $j=1,\ldots,k$ are finite, then*

$$n(\hat{\beta}_{\mathrm{OLS}} - \hat{\beta}_{\mathrm{WLS}})'\hat{\psi}^{-1}(\hat{\beta}_{\mathrm{OLS}} - \hat{\beta}_{\mathrm{WLS}}) \xrightarrow{\mathcal{L}} \chi_k^2$$

where

$$\hat{\psi} = (X'X/n)^{-1}\hat{V}_{\mathrm{OLS}}(X'X/n)^{-1} + (X'\Omega^{-1}X/n)^{-1}\hat{V}_{\mathrm{WLS}}(X'\Omega^{-1}X/n)^{-1}$$
$$- (X'X/n)^{-1}\hat{U}(X'\Omega^{-1}X/n)^{-1} - (X'\Omega^{-1}X/n)^{-1}\hat{U}(X'X/n)^{-1}$$

$$\hat{V}_{\mathrm{WLS}} = n^{-1}\sum_{i=1}^{n} W(Z_i)^2(Y_i - X_i\hat{\beta}_{\mathrm{WLS}})^2 X_i'X_t$$

$$\hat{U} = n^{-1}\sum_{i=1}^{n} W(Z_i)(Y_i - X_i\hat{\beta}_{\mathrm{OLS}})(Y_i - X_i\hat{\beta}_{\mathrm{WLS}})X_i'X_i$$

provided a.s. $\lim\hat{\psi}$ *is finite and nonsingular.*

PROOF. If $g(Z_i) \equiv X_i\beta_0$, we may write

$$\sqrt{n}(\hat{\beta}_{\mathrm{OLS}} - \hat{\beta}_{\mathrm{WLS}})$$
$$= (X'X/n)^{-1}n^{-1/2}\sum_{i=1}^{n} X_i'\varepsilon_i - (X'\Omega^{-1}X/n)^{-1}n^{-1/2}\sum_{i=1}^{n} W(Z_i)X_i'\varepsilon_i.$$

Now $(X'X/n)^{-1} \xrightarrow{\text{a.s.}} M_{XX}^{-1}$ and $(X'\Omega^{-1}X/n) \xrightarrow{\text{a.s.}} M_{\tilde{X}\tilde{X}}^{-1} \equiv E(W(Z_i)X_i'X_i)$ by Komolgorov's strong law of large numbers under A1, A3–A5. Hence

$$\sqrt{n}(\hat{\beta}_{\mathrm{OLS}} - \hat{\beta}_{\mathrm{WLS}}) \xrightarrow{L} M_{XX}^{-1}n^{-1/2}\sum_{i=1}^{n} X_i'\varepsilon_i - M_{\tilde{X}\tilde{X}}^{-1}n^{-1/2}\sum_{i=1}^{n} W(Z_i)X_i'\varepsilon_i$$

$$= n^{-1/2}\sum_{i=1}^{n} M_{XX}^{-1}X_i'\varepsilon_i - M_{\tilde{X}\tilde{X}}^{-1}W(Z_i)X_i'\varepsilon_i.$$

The random vectors $M_{XX}^{-1}X_i'\varepsilon_i - M_{\tilde{X}\tilde{X}}^{-1}W(Z_i)X_i'\varepsilon_i$ are i.i.d. with expectation zero given A4 and A5 and have covariance matrix

$$\psi = E([M_{XX}^{-1}X_i'\varepsilon_i - M_{\tilde{X}\tilde{X}}^{-1}W(Z_i)X_i'\varepsilon_i][\varepsilon_i X_i M_{XX}^{-1} - W(Z_i)\varepsilon_i X_i M_{\tilde{X}\tilde{X}}^{-1}])$$
$$= M_{XX}^{-1}V(\beta_0)M_{XX}^{-1} + M_{\tilde{X}\tilde{X}}\tilde{V}(\beta_0)M_{\tilde{X}\tilde{X}}^{-1} - M_{XX}^{-1}U(\beta_0)M_{\tilde{X}\tilde{X}}^{-1} - M_{\tilde{X}\tilde{X}}U(\beta_0)M_{XX}^{-1}$$

where

$$\tilde{V}(\beta) = E(W(Z_i)^2(Y_i - X_i\beta)^2 X_i'X_i)$$

$$U(\beta) = E(W(Z_i)(Y_i - X_i\beta)^2 X_i'X_i).$$

It follows by the multivariate Lindeberg-Levy central limit theorem that

$$\sqrt{n}(\hat{\beta}_{\mathrm{OLS}} - \hat{\beta}_{\mathrm{WLS}}) \xrightarrow{A} N(0, \psi)$$

provided ψ is finite. If ψ is nonsingular then

$$\sqrt{n}(\hat{\beta}_{\mathrm{OLS}} - \hat{\beta}_{\mathrm{WLS}})'\psi^{-1}(\hat{\beta}_{\mathrm{OLS}} - \hat{\beta}_{\mathrm{WLS}}) \xrightarrow{A} \chi_k^2$$

and the result follows immediately if $\hat{\psi}^{-1} \xrightarrow{p} \psi^{-1}$. In fact, $\hat{\psi}^{-1} \xrightarrow{\text{a.s.}} \psi^{-1}$. This follows since $(X'X/n)^{-1} \xrightarrow{\text{a.s.}} M_{XX}^{-1}$, $(X'\Omega^{-1}X/n)^{-1} \xrightarrow{\text{a.s.}} M_{\bar{X}\bar{X}}^{-1}$ as shown above, since $\hat{V}_{\mathrm{OLS}} \xrightarrow{\text{a.s.}} V(\beta_0)$ as shown in Theorem 3, and since $\hat{V}_{\mathrm{WLS}} \xrightarrow{\text{a.s.}} \bar{V}(\beta_0)$ and $\hat{U} \xrightarrow{\text{a.s.}} U(\beta_0)$ under the additional condition A5 by arguments identical to those proving $\hat{V}_{\mathrm{OLS}} \xrightarrow{\text{a.s.}} V(\beta_0)$ in Theorem 3. These facts imply $\hat{\psi} \xrightarrow{\text{a.s.}} \psi$, and arguments identical to those of Theorem 3 guarantee the finiteness of ψ. The nonsingularity of ψ implies the existence of $\hat{\psi}^{-1}$ almost surely for n sufficiently large; it follows that $\hat{\psi}^{-1} \xrightarrow{\text{a.s.}} \psi^{-1}$, which implies

$$\sqrt{n}(\hat{\beta}_{\mathrm{OLS}} - \hat{\beta}_{\mathrm{WLS}})\hat{\psi}^{-1}(\hat{\beta}_{\mathrm{OLS}} - \hat{\beta}_{\mathrm{WLS}}) \xrightarrow{A} \chi_k^2. \quad \square$$

THEOREM 5. *If* A1, A3, A4 *and* A6 *hold, then* $\hat{\beta}_{\mathrm{WLS}} \xrightarrow{\text{a.s.}} \beta^*$.

PROOF. We may write

$$\hat{\beta}_{\mathrm{WLS}} = \left(n^{-1} \sum_{i=1}^{n} W_i X_i' X_i\right)^{-1}\left(n^{-1} \sum_{i=1}^{n} W_i X_i' Y_i\right).$$

By Komolgorov's strong law of large numbers

$$\left(n^{-1} \sum_{i=1}^{n} W_i X_i' X_i\right) \xrightarrow{\text{a.s.}} E(W_i X_i' X_i)$$

and

$$\left(n^{-1} \sum_{n=1}^{n} W_i X_i' Y_i\right) \xrightarrow{\text{a.s.}} E(W_i X_i' Y_i)$$

given A1, A3, A4 and A6. Now $E(W_i)$ is finite,

$$E(W_i X_i' X_i) = E(W_i) M_{XX}$$

and

$$\begin{aligned} E(W_i X_i' Y_i) &= E(W_i) M_{Xg} + E(W_i X_i' \varepsilon_i) \\ &= E(W_i) M_{Xg} \end{aligned}$$

given A6. Since $E(W_i) > \delta$, we have

$$\begin{aligned} \hat{\beta}_{\mathrm{WLS}} &\xrightarrow{\text{a.s.}} E(W_i)^{-1} M_{XX}^{-1} E(W_i) M_{Xg} \\ &= M_{XX}^{-1} M_{Xg} = \beta^*. \quad \square \end{aligned}$$

LEMMA 1. *Suppose q statistical tests are performed, and let the event A_i denote accepting hypothesis H_0 on the basis of the i-th test. Let* $\bigcap_{i=1}^{q} A_i \subset B$, *the* *event denoting acceptance of hypothesis H_0 on the basis of the q tests taken together (i.e., the null hypothesis is accepted on the basis of q tests if each test accepts the null hypothesis). Then the size of the overall test, $P[\bar{B} \mid H_0]$, is such that*

$$P[\bar{B} \mid H_0] \leq \sum_{i=1}^{q} P[\bar{A}_i \mid H_0]$$

In particular, if $P[\bar{A}_i \mid H_0] = \alpha$, $i = 1, \ldots, q$, then $P[\bar{B} \mid H_0] \leq q\alpha$.

PROOF. The event $\bigcap_{i=1}^{q} A_i \mid H_0$ implies $B \mid H_0$. Then, by the implication rule Lukacs [1975, p. 7])

$$P[\bar{B} \mid H_0] \leq \sum_{i=1}^{q} P[\bar{A}_i \mid H_0]. \quad \square$$

REFERENCES

AGHEVLI, BIJAN B. AND MOSHEN S. KHAN, "Inflationary Finance and the Dynamics of Inflation: Indonesia 1951–1972," *American Economic Review*, 67 (June, 1977), 390–403.

ATKINSON, S. E. AND ROBERT HALVORSEN, "Interfuel Substitution in Steam Electric Power Generation," *Journal of Political Economy*, 84 (October, 1976), 959–978.

BARTLE, ROBERT, *The Elements of Integration* (New York: John Wiley and Sons, 1966).

CHRISTENSEN, L. R., D. W. JORGENSON AND L. J. LAU, "Transcendental Logarithmic Production Frontiers," *Review of Economics and Statistics*, 55 (February, 1973), 28–45.

CRAMÉR, HARALD, *Mathematical Methods of Statistics* (Princeton: Princeton University Press, 1946).

CRAMER, J., *Empirical Econometrics* (London: North Holland Publishing Co., 1969).

DENNY, MICHAEL AND MELVYN FUSS, "The Use of Approximation Analysis to Test for Separability and the Existence of Consistent Aggregates," *American Economic Review*, 67 (June, 1977), 404–418.

GLEJSER, H., "A New Test for Heteroskedasticity," *Journal of the American Statistical Association*, 64 (March, 1969), 316–323.

GRIFFIN, JAMES M., "Joint Production Technology: The Case of Petrochemicals," *Econometrica*, 46 (March, 1978), 379–398.

HAUSMAN, J. A., "Specification Tests in Econometrics," *Econometrica*, 46 (November, 1978) 1251–1272

KLOEK, T., "Note on a Large-Sample Result in Specification Analysis," *Econometrica*, 43 (December, 1975), 933–936.

LUKACS, EUGENE, *Stochastic Convergence* (New York: Academic Press, 1975).

MINCER, JACOB, *Schooling, Experience and Earnings* (New York: Columbia University Press, 1974).

MYERS, RAYMOND H. AND S. J. LAHODA, "A Generalization of the Response Surface Mean Square Error Criterion with a Specific Application to the Slope," *Technometrics*, 17 (November, 1975), 481–486.

PRAIS, S. J. AND H. S. HOUTHAKKER, *The Analysis of Family Budgets* (Cambridge: Cambridge University Press, 1971).

RICE, JOHN, *The Approximation of Functions* (Reading, Mass.: Addison Wesley Publishing Co., 1969).

SPADY, R. H. AND A. F. FRIEDLANDER, "Hedonic Cost Functions for the Regulated Trucking Industry," *The Bell Journal of Economics*, 9 (Spring, 1978), 159–179.

THEIL, HENRI, *Principles of Econometrics* (New York: John Wiley and Sons, 1971).

WALES, TERENCE J., "On the Flexibility of Flexible Functional Forms: An Empirical Approach," *Journal of Econometrics*, 5 (March, 1977), 183–193.

WHITE, HALBERT, "Using Nonlinear Least Squares to Approximate Unknown Regression Functions," University of Rochester Department of Economics Discussion Paper 77–3 (revised, August, 1979).

——— AND LAWRENCE OLSON, "Determinants of Wage Change on the Job: A Symmetric Test of Non-nested Hypotheses," University of Rochester Department of Economics Discussion Paper 77–17 (revised, September, 1977).