ETLA ELINKEINOELÄMÄN TUTKIMUSLAITOS
THE RESEARCH INSTITUTE OF THE FINNISH ECONOMY
Kalevankatu 3 B, 00100 Helsinki 10, Finland, tel. 601322

Keskusteluaiheita
Discussion papers

Timo Teräsvirta

SUPERIORITY COMPARISONS
OF HETEROGENEOUS LINEAR
ESTIMATORS

No. 127 23 December 1982

SUPERIORITY COMPARISONS OF HETEROGENEOUS LINEAR ESTIMATORS

by

Timo Teräsvirta

Abstract. In this paper conditions for strong and weak superiority of a heterogeneous linear estimator over another is derived. The general results are applied to same special cases: in particular, two restricted least squares estimators are compared using the superiority conditions obtained. The weak superiority criterion is used as a basis in forming an optimal sequence of tests (Anderson, 1962) for searching for the correct length of lag and appropriate degree of polynomial in the estimation of polynomial distributed lag models.

1. Introduction

During the last few years, a variety of biased estimators have been proposed alongside the previous ones like the restricted least squares (RLS) estimator for estimating the parameter vector of the general linear model. The performance of these estimators has been compared to that of the ordinary least squares (OLS) estimator mostly by using superiority criteria based on quadratic risk. Among the first examples of such comparisons are the superiority condition for the RLS estimator to be superior to the OLS estimator (Toro-Vizcarrondo and Wallace, 1968) and the dominance results for James-Stein estimators, for discussion see e.g. Judge et al. (1980) and Vinod and Ullah (1981).

Fewer results have been available on comparisons between biased linear estimators. However, such comparisons have also been made in econometric literature: for instance the problem whether to omit unobservables or substitute proxy variables for them in linear models is equivalent to comparing two different biased estimators. Hocking et al. (1976) have compared certain homogeneous linear estimators with each other. More recently, Trenkler (1980) and Teräsvirta (1982) have compared general homogeneous linear estimators using the generalised mean square error as the superiority criterion. Teräsvirta (1981a) has in particular discussed the relationship between the mixed and minimax estimators on that basis, while Guilkey and Price (1981) have carried out comparisons between RLS estimators. Price (1982) has included certain types of homogeneous linear estimators in his comparisons but without a general framework.

In this paper, a general framework is set up for comparisons between heterogeneous linear estimators. The special cases discussed in the literature can thereafter be treated in a straightforward fashion by applying the general theorem. A few examples will be considered here and, as will be seen, not only for illustration. The plan of the paper is as follows: In Section 2 concepts for comparing heterogeneous linear estimators are defined. In the next section, a general theorem for establishing strong superiority of a biased linear estimator over another is constructed. Section 4 contains examples of the theorem: various shrinkage estimators are considered as well as the comparisons between two RLS estimators and two principal component estimators. The use of proxy variables is also discussed assuming fixed proxies. Section 5 discusses weak superiority of biased linear estimators. The examples include RLS estimators and, in particular, the polynomial distributed lag estimator which has been widely applied in empirical work. Proxy variables are considered in that context, too. The final section contains a brief summary of results.

2. Preliminaries

Assume a linear model

$$y = X\beta + \varepsilon, \ \varepsilon \sim N(0, \sigma^2 I) \tag{2.1}$$

where y and ε are $n \times 1$ stochastic vectors, X is an $n \times p$ fixed matrix with $\text{rank}(X) = p$, β is a $p \times 1$ vector of regression coefficients, and σ^2 is the error variance. Define two linear heterogenous estimators of β as $b_j = D_j y + h_j$, $= 1,2$, where D_j is a fixed $n \times p$ matrix and h_j a fixed $p \times 1$ vector. In this paper, the interest will be focussed upon the conditions under which one of these estimators is better than the other. For considering the problem and making comparisons, we have to define superiority of an estimator over another in quantitative terms. Following established practice we make use of quadratic risk functions and define the strong superiority of b_2 over b_1 (cf. also Toro-Vizcarrondon and Wallace, 1968) at a single point (β,σ^2) in the parameter space as follows:

Definition 1. *Estimator* b_2 *is strongly superior to* b_1 *at* (β,σ^2) *if and only if*

$$E(b_1 - \beta)'A(b_1 - \beta) \geq E(b_2 - \beta)'A(b_2 - \beta) \tag{2.2}$$

for all loss matrices $A \geq 0$.[1)]

This definition is equivalent to requiring that the difference of two MSE matrices

$MSE(b_1) - MSE(b_2) \geq 0$ where $MSE(b_j) = E(b_j - \beta)(b_j - \beta)'$, cf. Theobald (1974).

Less restrictive definitions for superiority have been constructed by relaxing the restriction that the inequality (2.2) has to be valid for all non-negative definite loss matrices simultaneously and choosing a single loss matrix A_0 instead. Popular and well-founded choices of A have been I and X'X, see e.g. Wallace (1972) and Judge et al. (1980, pp. 24-26). In this paper we use

Definition 2. *Estimator* b_2 *is weakly superior to* b_1 *at* (β,σ^2) *if and only if* $E(b_1 - \beta)'X'X(b_1 - \beta) \geq E(b_2 - \beta)'X'X(b_2 - \beta)$.

In this definition the loss is related to the performance of the estimator in predicting $X\beta$, the conditional expectation of y given X and β. We shall employ it because of some rather attractive practical results that can be derived using it rather than any other A_0. Wallace (1972) has called Definition 2 the second weak MSE criterion.

The expectations in Definition 2 are sometimes called predictive MSE's of b_1 and b_2. Both definitions stress the fact that the superiority is defined at a single point of the parameter space at a time. They can nevertheless be generalised to larger subsets, for one such generalisation with application see Teräsvirta (1983). In what follows we shall omit the mention "at (β,σ^2)" for brevity, but the superiority remains as defined above throughout the discussion.

3. Conditions for strong superiority

For the purposes of this paper, it is convenient to write the MSE matrix as a decomposition into covariance and bias:

$$MSE(b_j) = E(b_j - \beta)(b_j - \beta)' = \sigma^2 D_j D_j' + d_j d_j'$$

where

$$d_j = H_j\beta + h_j \quad \text{with} \quad H_j = D_j X - I, \; j = 1,2 \; .$$

Set $C = D_1 D_1' - D_2 D_2'$ so that

$$\Delta_{12} = MSE(b_1) - MSE(b_2) = \sigma^2 C + d_1 d_1' - d_2 d_2' \; . \tag{3.1}$$

As pointed out above, b_2 is strongly superior to b_1 if and only if (3.1) is non-negative definite. Assume that we have the following decomposition

$$C = KLK', \; d_j = Kf_j, \quad j = 1,2 \tag{3.2}$$

where K is p x r, $r \leq p$, L is r x r, and f_j = r x 1, j = 1,2. This decomposition is useful whenever we want to compare estimators with singular covariance matrices. The difference (3.1) can now be written as

$$\Delta_{12} = K(\sigma^2 L + f_1 f_1' - f_2 f_2')K' \; .$$

It is well-known that $\Delta_{12} \geq 0$ if and only if

$$\sigma^2 L + f_1 f_1' - f_2 f_2' \geq 0. \tag{3.3}$$

Let us first exclude the trivial possibility that $L \geq 0$ and $f_2 = \alpha f_1$, $|\alpha| < 1$. This means that we do not consider any estimator b_2 with both smaller variance and bias than b_1, since this is a very rare case in practice. However, the assumption $L \geq 0$ is retained as yet. For (3.3) to hold it is then necessary that $\sigma^2 L + f_1 f_1' > 0$. This last assumption implies that either rank(L) = r-1 and f_1 is linearly independent of the columns of L or that $L > 0$. In both cases, (3.3) is equivalent to the following condition, cf. Farebrother (1976),

$$f_2'(\sigma^2 L + f_1 f_1')^{-1} f_2 \leq 1. \tag{3.4}$$

If we can assert that $L > 0$ then, using a well-known matrix identity, (3.4) can be written in the form

$$\sigma^{-2}\{f_{22} - f_{21}^2(\sigma^2 + f_{11})^{-1}\} \leq 1 \tag{3.5}$$

where

$$f_{ij} = f_i' L^{-1} f_j \ , \ i,j = 1,2 \ .$$

This is the main result of this section. A corresponding condition for two homogeneous linear estimators when K = I and $L > 0$ is to be found in Teräsvirta (1982). From (3.5), a sufficient but generally not necessary condition for (3.3) to hold when $L > 0$ is

$$\sigma^{-2} f_{22} \leq 1 \tag{3.6}$$

see also Trenkler (1980). If b_1 is unbiased then $f_1 = 0$ and (3.6) is necessary as well.

Assume now that $L < 0$. A lemma in Guilkey and Price (1981) states that (3.3) can then be valid only if L is a scalar, i.e. if $r = 1$. Then $\Delta_{12} \geq 0$ if and only if, in obvious notation,

$$\sigma^2 l_{11} + f_1^2 - f_2^2 \geq 0. \tag{3.7}$$

Note that if $L > 0$, then

$$\sigma^2 L + f_1 f_1' - f_2 f_2' \leq 0 \tag{3.8}$$

is a necessary condition for b_1 to be strongly superior to b_2.

Reverse now the rôles of b_1 and b_2 in Definition 1, so that $-L > 0$ and (3.3) becomes

$$-\sigma^2 L + f_2 f_2' - f_1 f_1' \geq 0. \tag{3.9}$$

Then it is seen from (3.8) that if $r = 1$, the necessary condition is also sufficient while this is not so when $r > 1$.

The above results can be formulated as

Theorem 1. *Assume linear model (2.1) and two heterogeneous linear estimators* $b_j = D_j y + h_j$, $j = 1,2$. *Set* $C = D_1 D_1' - D_2 D_2'$, *assume decomposition (3.2) and furthermore that* $\sigma^2 L + f_1 f_1' > 0$. *Then* b_2 *is strongly superior to* b_1 *if and only if (3.4) holds. If it is assumed that* $L > 0$ *then the strong superiority is equivalent to (3.5). On the other hand, if* $L < 0$ *then* b_2 *is strongly superior to* b_1 *if and only if* L *is a scalar and (3.7) is valid.*

In practice, $L > 0$ (or $L < 0$) seems to be a standard situation while $L \geq 0$ combined with the rank and linear independence conditions remains a more theoretical possibility.

4. Examples

In this section we shall apply Theorem 1 to some special cases of both homogeneous and heterogeneous estimators previously discussed in the statistical and econometric literature.

4.1. Shrinkage estimator with fixed shrinkage factor

Let $b_1 = c_1 b$ and $b_2 = c_2 b$ where $b = UX'y$ with $U = (X'X)^{-1}$, and $0 < c_2 < c_1 \leq 1$ are two fixed constants (Mayer and Willke, 1973). Then $C = (c_1^2 - c_2^2)U > 0$ and $H_j = (c_j - 1)I$, $h_j \equiv 0$, $j = 1,2$. Setting $K = I$ so that $L = C$ we have from (3.5), after some manipulation, that b_2 is strongly superior to b_1 if and only if

$$\sigma^{-2} c\beta'U\beta \leq 1 \tag{4.1}$$

where

$$c = \{(1 - c_2)^2 - (1 - c_1)^2\}(c_1^2 - c_2^2).$$

This result is also given in Price (1982). Improving b_1 is thus only possible by lowering the shrinkage factor, i.e. increasing shrinkage. One obvious consequence is that the new estimator b_2 is superior to the OLS estimator in a smaller subset than b_1, for further discussion see Teräsvirta (1981b).

4.2. Mixed and ridge estimators

Assume that we use stochastic prior information

$$r = R\beta + \phi_1 \tag{4.2}$$

where r is an m x 1 stochastic vector, R is an m x p fixed matrix with rank $m \leq p$, and it is also assumed that $\phi_1 \sim N(0,(\sigma^2/k_1)I)$, $k_1 > 0$. Suppose that in reality this information is biased so that

$$Er = R\beta + s_1 \tag{4.3}$$

where $s_1 \neq 0$, see Theil and Goldberger (1961), Yancey et al. (1974) and Teräsvirta (1981c). Combining (4.2) with the sample information (2.1) yields the mixed estimator

$$b_R(k_1) = (X'X + k_1R'R)^{-1}(X'y + k_1R'r).$$

Compare this with another mixed estimator $b_R(k_2)$ where R and (4.3) are the same as above but the uncertainty of prior information is altered in such a way that ϕ_1 in (4.2) is replaced by $\phi_2 \sim N(0,(\sigma^2/k_2)I)$, $k_2 > 0$. To find out when $b_R(k_2)$ is strongly superior to $b_R(k_1)$, we need

$$C = UR'(S_{k_2} - S_{k_1})RU$$

where

$$S_{k_j} = (k_j^{-1}I + RUR')^{-1}, \quad j = 1,2$$

see Teräsvirta (1981c). As $d_j = UR'S_{k_j} s_j$, we can choose $K = UR'$. Since for two pd matrices A and B, $A - B > 0$ implies $B^{-1} - A^{-1} > 0$ we conclude that $L = S_{k_1} - S_{k_2} > 0$ if and only if $k_2 > k_1$. If $k_1 = k_2$ then $L = 0$. Thus we can improve upon $b_R(k_1)$ only by choosing $k_2 > k_1$ if $m > 1$. When $k_2 \to \infty$, $b_R(k_2)$ converges towards the restricted least squares (RLS) estimator b_R. Thus, for some combinations of X, β and σ^2, a mixed estimator can be improved upon by a RLS estimator.

For two minimax estimators (Kuks and Olman, 1972) $b_I^*(k_j) = (X'X + k_j R'R)^{-1} X'y$, $j = 1,2$, we have

$$\Delta_{12} = \sigma^2 UR'(S_{k_2} T_{k_2}^{-1} S_{k_2} - S_{k_1} T_{k_1}^{-1} S_{k_1})RU + UR'S_{k_1}\beta_1\beta_1'S_{k_1}RU - UR'S_{k_2}\beta_2\beta_2'S_{k_2}RU \quad (4.3)$$

where

$$T_{k_j} = (2k_j^{-1}I + RUR')^{-1}, \quad j = 1,2.$$

Matrices S_{k_1}, S_{k_2}, T_{k_1} and T_{k_2} have the same eigenvectors, and since the eigenvalues of $S_k T_k^{-1} S_k$ are monotonously increasing functions of k, it follows that in (4.3), $S_{k_2} T_{k_2}^{-1} S_{k_2} - S_{k_1} T_{k_1}^{-1} S_{k_1} > 0$ if and only if $k_2 > k_1$. When $R = I$, the minimax estimator is simply the ridge estimator. Thus if $p > 1$, a ridge estimator $b_I^*(k_1)$ with a fixed ridge parameter k_1 can in some parts of the parameter space be improved upon by increasing the value of the ridge parameter while this is not possible by decreasing the value.

4.3. Restricted least squares estimators

In the previous sub-section the set (Er, R) was kept unchanged throughout. Here we start from two separate sets of linear restrictions

$$\tilde{r}_j = \tilde{R}_j \beta, \qquad j = 1,2$$

leading to the RLS estimators

$$b_{\tilde{R}_j} = b + U\tilde{R}_j'(R_j U\tilde{R}_j')^{-1}\,\hat{\tilde{s}}_j, \quad j = 1,2 \tag{4.4}$$

where $\hat{\tilde{s}}_j = \tilde{r}_j - \tilde{R}_j b$. Note that $\tilde{r}_j$ is now deterministic. Our aim is to derive conditions for strong superiority of $b_{\tilde{R}}$ over $b_{\tilde{R}_1}$ at (β, σ^2). Following Guilkey and Price (1981) we define

$$\tilde{R}_j = \begin{bmatrix} R_j \\ R_3 \end{bmatrix}, \quad \tilde{r}_j = \begin{bmatrix} r_j \\ r_3 \end{bmatrix}, \quad \tilde{s}_j = \begin{bmatrix} s_j \\ s_3 \end{bmatrix} = \begin{bmatrix} r_j - R_j\beta \\ r_3 - R_3\beta \end{bmatrix}, \quad j = 1,2 \tag{4.5}$$

so that $r_3 = R_3\beta$ is a common subset of restrictions for both sets. Let r_j be an $m_j \times 1$ vector, and R_j an $m_j \times p$ matrix, $1 < m_1 + m_2 + m_3 < p$, $j = 1,2$. Let $m_j = 0$ symbolise the absence of the j^{th} set of restrictions. We have

$$C = U\tilde{R}_2'\,(\tilde{R}_2 U\tilde{R}_2')^{-1}\tilde{R}_2 U \;-\; U\tilde{R}_1'\,(\tilde{R}_1 U\tilde{R}_1')^{-1}\tilde{R}_1 U \tag{4.6}$$

and

$$d_j = U\,\tilde{R}_j'(\tilde{R}_j U\tilde{R}_j')^{-1}\,\tilde{s}_j, \quad j = 1,2.$$

Using (4.5) we can write

$$U\tilde{R}_j'(\tilde{R}_jU\tilde{R}_j')^{-1}\,\tilde{R}_jU = UR_3'\,(R_3UR_3')^{-1}\,R_3U + UB\,R_j'\,D_{jj\cdot 3}^{-1}\,R_j\,B'U,\ j=1,2 \quad (4.7)$$

where

$$D_{jj\cdot 3} = R_jUR_j' - R_jUR_3'(R_3UR_3')^{-1}\,R_3UR_j',\quad j=1,2$$

and

$$B = I - R_3'\,(R_3UR_3')^{-1}\,R_3U.$$

The first term on the r.h.s. of (4.7) is the contribution of the restrictions $r_3 = R_3\beta$ to the covariance matrix of $b_{\tilde{R}_j}$, whereas the second term represents the remaining contribution of $r_j = R_j\beta$ after purging out the effect of $r_3 = R_3\beta$. Making use of (4.7) in (4.6) yields

$$C = UB(R_2'D_{22\cdot 3}^{-1}\,R_2 - R_1'\,D_{11\cdot 3}^{-1}\,R_1)\,B'U\ . \quad (4.8)$$

The matrix in parentheses in (4.8) is indefinite. Conforming to the block division in (4.5) we also have

$$d_j = U\{BR_j'D_{jj\cdot 3}^{-1}s_j + (I - BR_j'D_{jj\cdot 3}^{-1}R_jU)R_3'(R_3UR_3')^{-1}s_3\},\quad j=1,2. \quad (4.9)$$

From (4.8) we see at once that Theorem 1 does not apply as a proper L cannot be found.

Assuming $m_1 = 0$ so that $\tilde{R}_1 = R_3$ makes C positive semidefinite, but yet no decomposition with rank(L) $\geq$ p-1 exists since $m_2 + m_3 < p$. Only if we set $s_3 = 0$, thus supposing that the common restrictions $r_3 = R_3\beta$ are true, do we make progress and can choose $K = UBR_2'D_{22.3}^{-1}$ since $d_1 = 0$. Applying Theorem 1 we can conjecture that

$$\Delta_{12} = UBR_2'D_{22.3}^{-1}(\sigma^2 D_{22.3} - s_2 s_2')D_{22.3}^{-1}R_2B'U \geq 0$$

if and only if

$$\sigma^{-2}s_2'D_{22.3}^{-1}s_2 \leq 1. \tag{4.10}$$

Note that $s_3 = 0$ together with $m_1 = 0$ make $b_{\tilde{R}_1}$ unbiased. If we compare two biased RLS estimators, no condition for strong superiority of one over the other can be established in the general case. If $m_3 = 0$, (4.10) collapses into the condition of Toro-Vizcarrondo and Wallace (1968) for the strong superiority of b_{R_2} over b. We have proved

Corollary 1. *Assume linear model* (2.1) *and two restricted least squares estimators* $b_{\tilde{R}_1}$ *and* $b_{\tilde{R}_2}$ *defined as in* (4.4). *Then* $b_{\tilde{R}_2}$ *is strongly superior to* $b_{\tilde{R}_1}$ *if and only if*

(i) *the restrictions* $\tilde{r}_1 = \tilde{R}_1\beta$ *are a subset in* $\tilde{r}_2 = \tilde{R}_2\beta$ *and they are true, and*

(ii) *inequality* (4.10) *holds.*

On the other hand, $b_{\tilde{R}_1}$ *is strongly superior to* $b_{\tilde{R}_2}$ *if* $m_1 = 1$, $m_2 = m_3 = 0$ ($b_{\tilde{R}_2} = b$) *and* (4.10) *is invalid.*

The result says in effect that the only possibility for improving a RLS estimator using linear restrictions is to incorporate further restrictions into the model. However, for any reduction in matrix risk, the original restrictions have to be unbiased. The above mentioned paper by Guilkey and Price (1981, Theorem 3) also contains results on comparing two RLS estimators, but in Corollary 1 they appear in the correct form.

Note that (4.10) is a testable hypothesis. Under (4.10) and conditionally on $s_3 = 0$, the statistic

$$F = \hat{\sigma}^{-2} m_2^{-1} \hat{s}_2' D_{22.3}^{-1} \hat{s}_2$$

where $\hat{\sigma}^2 = (n - p)^{-1} y'(I - XUX')y$ and $\hat{s}_2 = r_2 - R_2 b$, follows a non-central F distribution with m_2 and $n - p$ degrees of freedom and non-centrality parameter 1/2, cf. also Toro-Vizcarrondo and Wallace (1968).

The principal component estimator is a special case of the RLS estimator, see e.g. Judge et al. (1980, pp. 468-471). In this case the data-specific linear restrictions imposed on the model can never be exactly valid as the eigenvalues of X'X have been assumed positive. Every principal component estimator is thus biased, and we have

Corollary 2. *Assume model (2.1) and two principal component estimators. They are biased and none of them is strongly superior to the other. The OLS estimator is strongly superior to the principal component estimator if and only if exactly one principal component is omitted in the latter and (4.10) with* $m_3 = 0$ *does not hold.*

The corresponding result in Price (1982), based on Theorem 3 in Guilkey and Price (1981), is incorrect.

4.4. Proxy variables

In econometric modelling it sometimes occurs that not all the variables suggested by economic theory can be observed or collecting the data may be too costly. In empirical work the unobservables have either been replaced by proxy variables thought to be related to the unobservable phenomenon or omitted completely. The proxies have been taken to be either unobservables measured with stochastic error and thereby observable, see for instance McCallum (1972), Wickens (1972), Aigner (1974) and Maddala (1977), or fixed variables as in Frost (1979) and Ohtani (1981). Maddala (1977) discusses the principal differences between these two approaches.

In the spirit of our general result, only fixed proxy variables will be considered here. Write (2.1) as

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

where X_j is a $p_j \times 1$ matrix and rank $(X_j) = p_j$, and β_j is a $p_j \times 1$ vector of regression coefficients, respectively. Let Z_2 be the matrix of the fixed proxy variables replacing the unobservable X_2. Furthermore, let

$$b_1 = \begin{bmatrix} b_{11} \\ 0 \end{bmatrix} = \begin{bmatrix} U_1 X_1' y \\ 0 \end{bmatrix}$$

where $U_1 = (X_1'X_1)^{-1}$, be the OLS estimator of $\beta = (\beta_1' \ \beta_2')'$ when X_2 is

omitted, while

$$b_2 = \begin{bmatrix} b_{21} \\ b_{22} \end{bmatrix} = \begin{bmatrix} X_1'X_1 & X_1'Z_2 \\ Z_2'X_1 & Z_2'Z_2 \end{bmatrix}^{-1} \begin{bmatrix} X_1' \\ Z_2' \end{bmatrix} y \tag{4.11}$$

is the corresponding estimator based on the OLS estimation of

$$y = X_1\beta_1 + Z_2\beta_2 + \varepsilon^*.$$

Conforming to the above block division we can write

$$D_1D_1' = \begin{bmatrix} U_1 & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad D_2D_2' = \begin{bmatrix} U_1 + U_1X_1'Z_2F^{-1}Z_2'X_1U_1 & -F^{-1}Z_2'X_1U_1 \\ -U_1X_1'Z_2F^{-1} & F^{-1} \end{bmatrix}$$

where

$$F = Z_2'M_1Z_2 > 0 \text{ with } M_1 = I - X_1U_1X_1' .$$

Thus

$$-C = \begin{bmatrix} -U_1X_1'Z_2 \\ I \end{bmatrix} F^{-1} \begin{pmatrix} -Z_2'X_1U_1 & I \end{pmatrix} \geq 0. \tag{4.12}$$

In the errors-in-variables case with two regressors and one proxy, McCallum (1972) and Wickens (1972) have recommended the use of the proxy variable because it always lowers the bias of the estimator of the scalar β_1.

Here it can be noticed, see (4.12), that when the proxy variables are assumed fixed the variance of any of the components of b_2 is never smaller than the variance of the corresponding component of b_1. Emphasizing the importance of variance and neglecting bias would therefore lead us to recommend the omission of X_2 and never the use of proxy variables, a conclusion diametrically opposite to that of McCallum and Wickens.

For a more balanced analysis, it can be found from (4.12) that the only strong condition we can hope to derive in the general case is for the superiority of the "omission alternative" b_1 over the "proxy alternative" b_2.

But then,

$$d_1 = \begin{bmatrix} U_1 X_1' X_2 \\ - I \end{bmatrix} \beta_2 , \qquad d_2 = \begin{bmatrix} U_1 X_1' (I - Z_2 F^{-1} Z_2' M_1) \\ F^{-1} Z_2' M_1 \end{bmatrix} X_2 \beta_2 . \qquad (4.13)$$

Combining (4.12) and (4.13) it is obvious that Theorem 1 is not applicable and no strong superiority condition can be established. This is in accordance with the conclusion of Frost (1979) arrived at by different considerations.

The outcome is not altered if we only consider loss matrices $A = \mathrm{diag}(A_{11}, 0)$, $A_{11} \geq 0$ is $p_1 \times p_1$, i.e., restrict ourselves to the estimation of β_1 like e.g. McCallum (1972), Wickens (1972) and Aigner (1974). However, assume that X_1 and X_2 are in fact orthogonal. Then $\mathrm{cov}(b_{11}) - \mathrm{cov}(b_{21}) < 0$ as before while b_{11} is

unbiased. Thus, substituting proxy variables for X_2 both increases variance and introduces bias so that in that particular situation the omission of X_2 is always a better alternative that the proxy variables Z_2 when strong superiority is concerned.

5. Conditions for weak superiority

If a weak superiority condition is used, the main difference as compared to previous sections is that the condition will be a scalar condition. When the necessary and sufficient condition for weak superiority of b_2 over b_1 at (β,σ^2) does not hold, this indicates weak superiority of b_1 over b_2. This was generally not true for strong superiority.

Definition 2 for weak superiority of b_2 over b_1 yields, using (3.1),

$$\sigma^2 \mathrm{tr}(D_1'X'XD_1 - D_2'X'XD_2) = \sigma^2 \mathrm{tr}X'XC \geq d_2'X'Xd_2 - d_1'X'Xd_1 \ . \qquad (5.1)$$

In what follows we shall study some interesting special cases of (5.1).

5.1. Shrinkage estimator with fixed shrinkage factor

The two estimators to be compared are $b_1 = c_1 b$ and $b_2 = c_2 b$ with $0 < c_2 < c_1 \leq 1$. Then (5.1) becomes

$$\sigma^{-2}\{(c_2 - 1)^2 - (c_1 - 1)^2\}\beta'X'X\beta \leq (c_1^2 - c_2^2)p$$

or

$$\sigma^{-2}c\beta'X'X\beta \leq p \tag{5.2}$$

where

$$c = 2(c_1 + c_2)^{-1} - 1 .$$

As in the case of strong superiority, $c_1 > c_2$ is a necessary condition for the superiority of b_2 over b_1. Thus b_1 can sometimes be improved upon by shrinking further towards the origin. Note that (5.2) is a testable hypothesis. Under (5.2), $F = \hat{\sigma}^{-2}p^{-1}b'X'Xb$ with $\hat{\sigma}^2 = (n - p)^{-1}(y - Xb)'(y - Xb)$ follows a non-central F distribution with p and n - p degrees of freedom and non-centrality parameter p/2c. Rejecting (5.2) happens at large values of F and means preferring b_1 to b_2. Choosing $b_1 = b$ ($c_1 = 1$) we have a simple pre-test estimator which is a weighted combination of b and b_2. When $c_1 = 1$, and c_2 is very close to unity, (5.2) is rather likely to hold at a given point

(β,σ^2) even if there are always points in the parameter space where (5.2) is not valid. This conclusion is obvious from Teräsvirta (1981b), but here it is reached from a different viewpoint.

5.2. Mixed and ridge estimators

When strong superiority of mixed estimators was considered it was noticed that to improve estimation as compared to the estimator $b_R(k_1)$ with fixed k_1 it was necessary choose a constant $k_2 > k_1$. A similar result was seen to apply to ridge estimators. This ceases to be true when weak superiority is used as a criterion. Nordberg (1982) has demonstrated that as a function of k, the mean square error of the ridge estimator may have more than one minimum in $(0,\infty)$, and the same is obviously true for the predictive MSE as well. While (5.1) gives the necessary and sufficient superiority condition, it does not allow us to draw any straightforward analytical conclusions about the relationship between k_1 and k_2.

5.3. Restricted least squares estimators

In Section 4.3 conditions for strong superiority of a RLS estimator over another were found to be rather strict. In the following we shall see how much more relaxed the conditions for weak superiority will be.

For the weak superiority of $b_{\tilde{R}_2}$ over $b_{\tilde{R}_1}$ we have from (5.1), assuming $m_3 = 0$ for simplicity, that

$$\sigma^{-2}\{s_2'(R_2UR_2')^{-1}s_2 - s_1'(R_1UR_1')^{-1}s_1\} \leq m_2 - m_1 . \tag{5.3}$$

Inequality (5.3) implies that a necessary condition for superiority is that the number of linear restrictions in b_{R_2} is at least as great as in b_{R_1}. If $m_1 = 0$, (5.3) is simply the second weak MSE condition of Wallace (1972).

By restricting the above assumptions somewhat, useful results can be obtained. Suppose that the second set of linear restrictions is nested in the first, i.e., $R_2\beta = r_2$ implies $R_1\beta = r_1$, and $m_1 < m_2$. Thus we have $R_1 = GR_2$ and $r_1 = Gr_2$ where G is a $m_1 \times m_2$ matrix with rank m_1, and the following theorem can be shown to be valid:

Theorem 2. *Assume the linear model* (2.1) *and two sets of linear restrictions* $R_j\beta = r_j$, *where* R_j *is* $m_j \times p$, $m_1 < m_2 < p$, $j = 1,2$, *with* $R_1 = GR_2$ *and* $r_1 = Gr_2$. *Then*

$$\tilde{F} = \hat{\sigma}^{-2}(m_2 - m_1)^{-1}(\hat{s}_2'(R_2UR_2')^{-1}\hat{s}_2 - \hat{s}_1'(R_1UR_1')^{-1}\hat{s}_1) \tag{5.4}$$

where

$$\hat{\sigma}^2 = (n - p)^{-1}y'(I - XUX')y$$

and

$$\hat{s}_1 = G\hat{s}_2, \quad \hat{s}_2 = R_2b - r_2$$

follows a non-central F *distribution with* $m_2 - m_1$ *and* $n - p$ *degrees of freedom and non-centrality parameter* $(2\sigma^2)^{-1}\{s_2'(R_2UR_2')^{-1}s_2 - s_1'(R_1UR_1')^{-1}s_1\}$. The proof is in the appendix.

From Theorem 2 it follows that (5.3) is a testable hypothesis and when it is assumed valid, (5.4) has a non-central $F(m_2 - m_1, n - p, (m_2 - m_1)/2)$ distribution. Next we shall discuss an application of this result.

5.4. Polynomial distributed lag estimation

Assume a finite distributed lag model (2.1) where now the columns of X are consecutive lags of the first column. In small sample situations a popular estimation scheme has been the polynomial distributed lag estimation in which it is assumed that the regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$, lie on a polynomial of order q (Almon, 1965). This assumption can be expressed as a set of linear constraints $R\beta = 0$ where the j^{th} row of the $(p - q)x(p + 1)$ matrix R is

$$(0,\ldots,0,1,-\binom{q+1}{1},\ldots,(-1)^j\binom{q+1}{j},\ldots,(-1)^{q+1},0,\ldots,0)$$

cf. Shiller (1973) and Teräsvirta (1976). Note that the number of columns in X is now $p + 1$. The number of zeros at the beginning of the row is $j - 1$ and at the end $p - q - j$. A practical problem is that both the lag length p and the proper order of the lag polynomial q (the polynomial assumption is usually at most approximately correct for any low value of q) are unknown parameters. Various approaches to determining p and q, or q given p, from data have been suggested in the literature, for an overview see Hendry et al. (1982).

The above theory can be used for testing hypotheses about the lag length and the order of polynomial. Let R_q denote the constraint matrix when the order of polynomial is q. Then $R_q = G_q R_{q-1}$ where G_q is a $(p - q) \times (p - q + 1)$ matrix whose j^{th} row is $(0,\ldots,0,1,-1,0,\ldots,0)$. The number of zeros at the beginning is $j - 1$ and at the end $p - q - j - 1$. More generally

$$R_{q_1} = GR_{q_2}, \quad 0 < q_2 < q_1 < p, \tag{5.5}$$

where $G = G_{q_1} G_{q_1-1} \ldots G_{q_2+1}$, and using (5.4) we can test (5.3) that lowering the order of polynomial from $q_1 = p - m_1 - 1$ to $q_2 = p - m_2 - 1$, $q_1 > q_2$, reduces the predictive MSE.

The pairwise comparisons made possible by the above theory do not seem very expedient in choosing the best combination of p and q. However, based upon (5.3) and (5.5), a sequential testing procedure with certain optimal properties can be constructed for determining those two parameters. This procedure amounts to testing a sequence of nested hypotheses, cf. e.g. Anderson (1962; 1971, pp. 270-276) and Mizon (1977). We shall now take a closer look at this proposal.

Suppose that the maximum lag is at most p and degree of polynomial not higher than q. Furthermore, set

$$\bar{Q}_{p-j} = \sigma^{-2} \bar{s}'_{p-j} (\bar{R}_{p-j} U \bar{R}'_{p-j})^{-1} \bar{s}_{p-j}$$

where

$$\bar{R}_{p-j} = (0\ I_j),\ I_j \text{ is the } j \times j \text{ identity matrix}$$

and

$$\bar{s}_{p-j} = \bar{R}_{p-j}\beta.$$

We start by testing the linear hypothesis that excluding the longest lag reduces the PMSE, or,

$$H_{p-1}: \bar{Q}_{p-1} \leq 1. \tag{5.6}$$

No other restrictions are considered as yet. Define

$$\hat{\bar{Q}}_{p-j} = \hat{\sigma}^{-2}\hat{\bar{s}}'_{p-j}(\bar{R}_{p-j}U\bar{R}'_{p-j})^{-1}\hat{\bar{s}}_{p-j}\ ,\ j \geq 1.$$

where

$$\hat{\sigma}^2 = (n - p - 1)^{-1}y'(I - XUX')y$$

and set $\bar{F}_{p-1} = \hat{\bar{Q}}_{p-1}$, $\bar{F}_{p-j} = \hat{\bar{Q}}_{p-j} - \hat{\bar{Q}}_{p-j+1}\ ,\ j > 1.$

Under (5.6), $\bar{F}_{p-1} \sim F(1, n - p - 1, 1/2)$, cf. also Wallace (1972). If H_{p-1} is accepted, we proceed by testing the weak superiority of $b_{\bar{R}_{p-2}}$ over $b_{\bar{R}_{p-1}}$:

$$H_{p-2}|H_{p-1}: \bar{Q}_{p-2} - \bar{Q}_{p-1} \leq 1. \tag{5.7}$$

When (5.7) holds conditionally, $\bar{F}_{p-2}$ has the same non-central distribution as $\bar{F}_{p-1}$ under H_{p-1}. In general, after a sequence of accepted hypotheses $H_{p-1},\{H_{p-2}|H_{p-1}\},\ldots,\{H_{p-(j-1)}|H_{p-(j-2)},\ldots,H_{p-1}\}$ the next conditional hypothesis will be

$$H_{p-j}|H_{p-(j-1)},\ldots,H_{p-1}: \bar{Q}_{p-j} - \bar{Q}_{p-j+1} \leq 1. \tag{5.8}$$

Under (5.8), $\bar{F}_{p-j}$ has again an $F(1,n - p - 1, 1/2)$ distribution.

Assume now that the null hypothesis $\bar{Q}_{p-J-1} - \bar{Q}_{p-J} \leq 1$ is rejected. Then we move one step backward and compare the restricted least squares estimator with maximum lag p-J with the estimator with the same maximum lag and the polynomial restriction of order $q < p - J$. We have

$$\tilde{H}_q|H_{p-J},\ldots,H_{p-1}: \tilde{Q}_q - \bar{Q}_{p-J} \leq p - q \tag{5.9}$$

where $p - q \geq 1$, and

$$\tilde{Q}_q = \tilde{s}_q' (\tilde{R}_q U \tilde{R}_q')^{-1} \tilde{s}_q$$

with

$$\tilde{R}_q = \begin{bmatrix} R_q & 0 \\ 0 & I_J \end{bmatrix} \text{ and } \tilde{s}_q = \tilde{R}_q \beta .$$

Taking account of (5.9) and the fact that (5.8) for $j = J$ was accepted, we can test (5.9) conditionally on the earlier conditional

hypotheses. Under (5.9), in obvious notation, $\tilde{F}_q = \hat{\sigma}^{-2}(p-q)^{-1}(\hat{\tilde{Q}}_q - \hat{\tilde{Q}}_{p-J})$ follows an $F(p - q, n - p - 1, (p - q)/2)$ distribution because the sequence of linear hypotheses is nested. If (5.9) is accepted, the degree of polynomial is tentatively lowered by one, and the corresponding hypothesis is

$$\tilde{H}_{q-1} \Big| \tilde{H}_q, H_{p-J}, \ldots, H_{p-1} \colon \tilde{Q}_{q-1} - \tilde{Q}_q \leq 1. \tag{5.10}$$

Set $\tilde{F}_{q-j} = \hat{\tilde{Q}}_{q-j} - \hat{\tilde{Q}}_{q-j+1}$, $j \geq 1$. When (5.10) is valid, $\tilde{F}_{q-1}$ has a non-central $F(1, n - p - 1, 1/2)$ distribution. In general, after having accepted the previous hypotheses, we can test the weak superiority of $b_{\tilde{R}_{q-j}}$ over $b_{\tilde{R}_{q-j+1}}$, i.e.,

$$\tilde{H}_{q-j} \Big| \tilde{H}_{q-j+1}, \ldots, \tilde{H}_q, H_{p-J}, \ldots, H_{p-1} \colon \tilde{Q}_{q-j} - \tilde{Q}_{q-j+1} \leq 1. \tag{5.11}$$

Under (5.11), $\tilde{F}_{q-j} \sim F(1, n - p - 1, 1/2)$. When the first rejection occurs at degree $q - K - 1$, say, then the combination $(p - J, q - K)$ is chosen to represent the lag length and the proper order of the polynomial.

Another sequential procedure appears in Pagano and Hartley (1981), but there are dissimilarities between their method and the technique proposed here. Pagano and Hartley want to test whether the regression coefficients of longest lags are indeed zero and, thereafter, following Godfrey and Poskitt (1975) whether the polynomial restrictions are exactly true. In our approach less categorical

hypotheses are formulated and tested, but this is not crucial. The above hypotheses can be easily modified if desired. For instance, testing for the true lag length sounds reasonable since we have a finite lag model. But then, subsequent testing for the degree of the polynomial may well be carried out with the risk reduction in mind because any non-trivial polynomial restriction can hardly be expected to hold exactly in practice.

A more important difference is that our procedure is a direct generalisation of the optimal sequential testing procedure of Anderson (1962) to the PMSE reduction case while that is not true for the Pagano-Hartley procedure. The testing is carried out against the immediately preceding less restricted model and the test statistics are uncorrelated. If the null hypotheses (5.6), (5.7), (5.8), (5.9), (5.10) and (5.11) are altered by equating their left-hand sides to zero so that the corresponding test statistics follow central F distributions under the null hypotheses, the classical Anderson procedure is obtained, except for one dissimilarity. All our test statistics contain the same estimator $\hat{\sigma}^2$, whereas Anderson (1962) uses the residual variance estimated from the immediately preceding less restricted model to which the model to be tested is compared. There is, however, a natural explanation to this difference. If an hypothesis that a linear restriction is valid is accepted, then a conditional unbiased estimate for the error variance σ^2 is obtained from that restricted model and it can be used at the next stage. Since our procedure does not consider the truth of the restrictions but only their usefulness in reducing PMSE, the only unbiased estimator available is $\hat{\sigma}^2$.

On the other hand, when Pagano and Hartley (1981) search for the correct lag length they test the following sequence of hypotheses till the first rejection:

$$H^{*}_{p-j}: v'_{p-j}\beta = 0, \; j = 1,2,\ldots \tag{5.12}$$

where $v_{p-j} = (0,\ldots,0,v_{p-j+2,p-j+2},v_{p-j+2,p-j+3},\ldots,v_{p-j+2,p+1})'$. The number of zeros at the beginning is $p - j + 1$.

The test statistics corresponding to the sequence (5.12) are uncorrelated due to an orthogonal transformation of (2.1), but the problem is that this sequence of hypotheses does not correspond to that of the optimal procedure in the original parameter space because the hypotheses in (5.12) are not nested there. The same applies to the hypotheses concerning the degree of the polynomial. Incidentally, this leads the authors to employ $\hat{\sigma}^2$ as the estimator of σ^2 in all their test statistics. In short, the procedure of Pagano and Hartley (1981) is not optimal in the sense of Anderson (1962). Results upon the performance of the method suggested here will be reported in a forthcoming paper.

5.5. Proxy variables

In order to allow for proxy variables, we have to modify the definition of the predictive mean square error $E(\tilde{b} - \beta)'X'X(\tilde{b} - \beta)$ slightly.

Define the PMSE of the least squares estimator b_2 when proxy variables are employed as

$$PMSE(b_2) = E\{X_1(b_{21} - \beta_1) + Z_2(b_{22} - \beta_2)\}'\{X_1(b_{21} - \beta_1) + Z_2(b_{22} - \beta_2)\}$$

$$= \sigma^2(p_1 + p_2) + \beta_2' X_2 M_1 (I - Z_2 F^{-1} Z_2') M_1 X_2 \beta_2. \qquad (5.13)$$

Since

$$PMSE(b_1) = E(b_1 - \beta)'X'X(b_1 - \beta) = \sigma^2 p_1 + \beta_2' X_2' M_1 X_2 \beta_2$$

we have

$$PMSE(b_2) - PMSE(b_1) = \sigma^2 p_2 - \beta_2' X_2' M_1 Z_2 F^{-1} Z_2' M_1 X_2 \beta_2$$

so that a necessary and sufficient condition for b_1 to be weakly superior to b_2 is

$$\sigma^{-2} \beta_2' X_2' M_1 Z_2 F^{-1} Z_2' M_1 X_2 \beta_2 \leq p_2. \qquad (5.14)$$

By choosing $p_1 = p_2 = 1$, (5.14) reduces to the superiority condition given by Ohtani (1981).

If X_1 is orthogonal to X_2 but not to Z_2, then (5.14) has the form

$$\sigma^{-2}\beta_2'X_2'Z_2F^{-1}Z_2'X_2\beta_2 \leq p_2 .$$

Note that the corresponding condition for b_1 to be superior to the OLS estimator b is

$$\sigma^{-2}\beta_2'X_2'M_1X_2\beta_2 \leq p_2. \tag{5.15}$$

From (5.13) it is seen that

$$\beta_2'X_2'M_1Z_2F^{-1}Z_2'M_1X_2\beta_2 \leq \beta_2'X_2'M_1X_2\beta_2 .$$

Thus, if (5.15) holds so does (5.14), and we can conjecture

Theorem 3. *Let b_1 be the estimator of β obtained by omitting p_2 variables X_2. If this estimator is weakly superior to the OLS estimator b then it is weakly superior to any estimator b_2 in which a set of proxy variables Z_2 have been substituted for X_2.*

The theorem simply says that if the unobservable variables do not have prediction power then it is not worthwhile to substitute proxies for them.

6. Summary

In this paper we have derived conditions for strong and weak superiority of a heterogeneous linear estimator over another. In particular, it is demonstrated that in the case of RLS estimators the strong superiority is indeed a very strong requirement while the use of weak superiority opens possibilities for testing for the superiority of a RLS estimator over another in a relatively general case. These tests can be fruitfully applied to the polynomial distributed lag estimation.

For various shrinkage estimators the general result is that they can only be improved upon by increasing shrinkage. As to the use of proxy variables in linear models, no condition for strong superiority in either direction exists while a necessary and sufficient condition is readily available for weak superiority. It also turns out that if an omission of variables yields a smaller PMSE than possessed by the OLS estimator of the full model, then this "omission" estimator is weakly superior to any least squares estimator with proxy variables.

References

Aigner, D.J., 1974, MSE dominance of least squares with errors-of-observation, *Journal of Econometrics* 2, 365-372.

Almon, S., 1965, The distributed lag between capital appropriations and expenditures, *Econometrica* 33, 178-196.

Anderson, T.W., 1962, The choice of degree of a polynomial regression as a multiple decision problem, *Annals of Mathematical Statistics* 33, 255-265.

Anderson, T.W., 1971, *The statistical analysis of time series* (Wiley, New York).

Farebrother, R.W., 1976, Further results on the mean square error of ridge regression. *Journal of the Royal Statistical Society* B 38, 248-250.

Frost, P.A., 1979, Proxy variables and specification bias, *Review of Economics and Statistics* 61, 323-325.

Godfrey, L.G. and D.S. Poskitt, 1975, Testing the restrictions of the Almon lag technique, *Journal of the American Statistical Association* 70, 105-108.

Goldberger, A.S., 1964, *Econometric theory* (Wiley, New York).

Guilkey, D.K. and J.M. Price, 1981, On comparing restricted least squares estimators, *Journal of Econometrics* 15, 397-404.

Hendry, D.F., A.R. Pagan and J.D. Sargan, 1982, Dynamic specification, LSE Econometrics Programme, Discussion Paper No. A26.

Hocking, R.R., F.M. Speed and M.J. Lynn, 1976, A class of biased estimators in linear regression. *Technometrics* 18, 425-437.

Judge, G.G., W.E. Griffiths, R.C. Hill and T.-C. Lee, 1980, *Theory and practice of econometrics* (Wiley, New York).

Kuks, J. and V. Olman, 1972, Minimaksnaja linejnaja ocenka koefficientov regressii. Eesti NSV teaduste akadeemia toimetised 21, 66-72.

Maddala, G.S., 1977, *Econometrics* (McGraw-Hill, New York).

Mayer, L.S. and T.A. Willke, 1973, On biased estimation in linear models. *Technometrics* 15, 497-508.

McCallum, B.T., 1972, Relative asymptotic bias from errors of omission and measurement. *Econometrica* 40, 757-758.

Mizon, G.E., 1977, Inferential procedures in nonlinear models: An application in a UK industrial cross section study of factor substitution and returns to scale, *Econometrica* 45, 1221-1241.

Nordberg, L., 1982, A procedure for determination of a good ridge parameter in linear regression, *Communications in Statistics* B11, 285-309.

Ohtani, K., 1981, On the use of a proxy variable in prediction: An MSE comparison, *Review of Economics and Statistics* 63, 627-628.

Pagano, M. and M.J. Hartley, 1981, On fitting distributed lag models subject to polynomial restrictions, *Journal of Econometrics* 16, 171-198.

Price, J.M., 1982, Comparisons among regression estimators under the generalized mean square error criterion. *Communications in Statistics* A11, 1965-1984.

Rao, C.R., 1965, *Linear statistical inference and its applications* (Wiley, New York).

Shiller, R.J., 1973, A distributed lag estimator derived from smoothness priors, *Econometrica* 41, 775-788.

Teräsvirta, T., 1976, A note on bias in the Almon distributed lag estimator, *Econometrica* 44, 1317-1321.

Teräsvirta, T., 1981a, A comparison of mixed and mimimax estimators of linear models, *Communications in Statistics* A10, 1765-1778.

Teräsvirta, T., 1981b, Restricted superiority of a shrinkage estimator with a fixed shrinkage factor, Research Institute of the Finnish Economy, Discussion Paper No. 84.

Teräsvirta, T., 1981c, Some results on improving the least squares estimation of linear models by mixed estimation, *Scandinavian Journal of Statistics* 8, 33-38.

Teräsvirta, T., 1982, Superiority comparisons of homogeneous linear estimators, *Communications in Statistics* A11, 1595-1601.

Teräsvirta, T., 1983, Restricted superiority of linear homogeneous estimators over ordinary least squares, *Scandinavian Journal of Statistics* (forthcoming).

Theil, H. and A.S. Goldberger, 1961, On pure and mixed statistical estimation in economics, *International Economic Review* 2, 65-78.

Theobald, C.M., 1974, Generalizations of mean square error applied to ridge regression, *Journal of the Royal Statistical Society* B36, 103-106.

Toro-Vizcarrondo, C. and T.D. Wallace, 1968, A test of the mean square error criterion of restrictions in linear regression, *Journal of the American Statistical Association* 63, 558-572.

Trenkler, G., 1980, Generalized mean square error comparisons of biased regression estimators, *Communications in Statistics* A9, 1247-1259.

Vinod, H.D. and A. Ullah, 1981, *Recent advances in regression methods* (Dekker, New York).

Wallace, T.D., 1972, Weaker criteria and tests for linear restrictions in regression, *Econometrica* 40, 689-698.

Wickens, M.R., 1972, A note on the use of proxy variables, *Econometrica* 40, 759-761.

Yancey, T.A., G.G. Judge and M.E. Bock, 1974, A mean square error test when stochastic restrictions are used in regression, *Communications in Statistics* 3, 755-768.

Appendix

Proof of Theorem 2: Consider first the numerator of $\tilde{F}$ and define

$$Q_j = \varepsilon' A_j \varepsilon, \quad j = 1,2$$

where

$$A_j = XUR_j'(R_jUR_j')^{-1}R_jUX' \ .$$

Then, because $R_1 = GR_2$, we have $A_1A_2 = A_2A_1 = A_1$ so that $A_2 - A_1$ is idempotent, and rank $(A_2 - A_1) = tr(A_2 - A_1) = m_2 - m_1$. Then $\sigma^{-2}(Q_2 - Q_1) \sim \chi^2(m_2 - m_1)$, cf. Rao (1965, p. 150). From this it follows (Rao, 1965, p. 150) that $Q = \sigma^{-2}\{\hat{s}_2'(R_2UR_2')^{-1}\hat{s}_2 - \hat{s}_1'(R_1UR_1')^{-1}\hat{s}_1\}$ has a non-central χ^2 distribution with $m_2 - m_1$ degrees of freedom and non-centrality parameter $(1/2)EQ = (2\sigma^2)^{-1}\{s_2'(R_2UR_2')^{-1}s_2 - s_1'(R_1UR_1')^{-1}s_1\}$. Note that $(I - XUX')A_j = 0$, $j = 1,2$. Since normality was assumed this implies that $\hat{\sigma}^2$ is independent of the numerator and the result of the theorem follows.

Footnote (on page 3)

1) $A \geq 0$ means that the square matrix A is non-negative definite. If A is positive definite, notation $A > 0$ will be used.