1984

Double k-Class Shrinkage Estimators in Multiple Regression

R AL Carter

10419

ISSN: 0318-725X
ISBN: 0-7714-0561-8

RESEARCH REPORT 8412

DOUBLE k-CLASS SHRINKAGE ESTIMATORS IN MULTIPLE REGRESSION

R. A. L. Carter[1]

Double k-Class Shrinkage Estimators in Multiple Regression

R. A. L. Carter[1]
Department of Economics
University of Western Ontario
London Canada

1. Introduction

The method of least-squares (maximum likelihood) has traditionally been used to estimate the slope parameters of the classical linear model because of its best-linear-unbiased property. In recent years work begun by James and Stein (1961), extended by Sclove (1968) and generalized by Ullah and Ullah (1978) has led to a family of shrinkage estimators which are of interest because their risk is less than that of least-squares (LS). These estimators were originally intended to provide estimates of all the regression coefficients jointly. However, one can think of several other uses to which they might naturally be put. For example, Aigner and Judge (1977) have applied them to the estimation of individual coefficients. Alternatively, they might be used to estimate linear combinations of coefficients or to provide forecasts.

The aim of this paper is to investigate the risk of the double k-class (KK) estimators in each of these applications and, in each case, to compare it to the risk of LS. In each case the most desirable values are obtained for the two arbitrary constants, k_1 and k_2, which appear in the expression for KK.

2. The Model and Estimators

Consider the classical linear model

$$y = X\beta + \epsilon \tag{2.1}$$

where y is a vector of T observations on a dependent variable, X is a T • K

matrix of T observations on K fixed regressors, β is a vector of K parameters to be estimated and ϵ is a vector of T independent realizations of an unobservable normal error with mean zero and variance σ^2. Assume X is of rank K and that C is the orthonormal matrix of characteristic vectors of $X'X$. Then an alternative formulation of the model is

$$y = Z\alpha + \epsilon \tag{2.2}$$

where $Z = XC$, $\alpha = C'\beta$ and $Z'Z = \Lambda$, a K order diagonal matrix whose diagonal elements are the K characteristic roots of $X'X$. We will assume that the columns of X are arranged so that the smallest root of $X'X$, λ_1, is the first diagonal element of Λ and the largest root, λ_K, is the last.

The least-squares (LS) estimators of β and α are

$$b = (X'X)^{-1}X'y \tag{2.3}$$

and

$$a = \Lambda^{-1}Z'y = C'b \tag{2.4}$$

The risks of b and a are

$$E(b-\beta)'(b-\beta) = \sigma^2 \operatorname{tr}(X'X)^{-1} = \sigma^2 \operatorname{tr}\Lambda^{-1} \tag{2.5}$$

$$= E(a-\alpha)'(a-\alpha)$$

The vector of LS residuals is

$$e = y - Xb = y - Za \tag{2.6}$$

Shrinkage estimators are obtained by multiplying b (or a) by a scalar lying between 0 and 1 and chosen so as to make the risk of the product less than that of b and a. A general form of the shrinkage estimator is given by Ullah and Ullah (1978) as:

$$\hat{\beta} = \left[1 - \frac{k_1 e'e}{y'y - k_2 e'e}\right] b \tag{2.7}$$

or, alternatively, as

$$(2.8) \qquad \hat{\alpha} = \left[1 - \frac{k_1 e'e}{y'y - k_2 e'e}\right] a = C'\hat{\beta}$$

The values k_1 and k_2 are arbitrary scalars whose presence led the Ullahs to label $\hat{\beta}$, "the double k-class estimator" (KK).

Various members of the KK family arise from varying the values of k_1 and k_2. An important set of estimators is obtained by setting $k_2 = 1.0$. Then (2.7) and (2.8) give the Stein-rule estimators (Stein (1956), James and Stein (1961), Bock (1975), Judge and Bock (1976)) which have smaller risk than LS for values of k_1 in the range $0 < k_1 \leq \frac{2(d_1-2)}{n+2}$, where $d_1 = \lambda_1 \, tr\Lambda^{-1}$ and $n = T-K$.

Ullah and Ullah have derived the exact and approximate bias, mean-squared-error (MSE) matrix and risk for $\hat{\beta}$ when[2] $k_1 > 0$ and $0 \leq k_2 \leq 1$.

A disadvantage of the shrinkage estimators in the form given above is that they shrink b towards zero. However, researchers often have linear hypotheses that β is some non-zero vector; that is they have hypotheses of the form

$$(2.9) \qquad R\beta = r$$

where R is a known $J \cdot K$ matrix of rank J and r is a known J element vector. A value of β which is a solution to (2.9) is

$$(2.10) \qquad \beta_o = R^{+}r = R'(RR')^{-1}r \, ,$$

where R^{+} is the Moore-Penrose inverse of R. Sometimes $J = K$, so that $R^{+} = R^{-1}$, or $R = I$ so that r is the hypothesized value of β. In the presence of hypotheses like (2.9) it is desirable to shrink[3] b towards β_o rather than towards zero. This leads to the estimators (see Appendix A for details)

$$(2.11) \qquad \hat{\beta} = (1-k_1 c)(b-\beta_o) + \beta_o$$

or

$$(2.12) \qquad \hat{\alpha} = (1-k_1 c)(a-\alpha_o) + \alpha_o$$

where $\alpha_o = C'\beta_o$,

$$(2.13) \qquad c = \frac{(y-X\beta_o)'M(y-X\beta_o)}{(y-X\beta_o)'(y-X\beta_o) - k_2(y-X\beta_o)'M(y-X\beta_o)}$$

and $M = I - X(X'X)^{-1}X'$. The special cases (2.7) and (2.8) are obtained from (2.11) and (2.12) by setting $\beta_o = \alpha_o = 0$.

3. Asymptotic Expansions

The exact bias, MSE and risk of $\hat{\beta}$, given in Appendix A, are inconvenient in that they do not suggest optimum values of the arbitrary constants k_1 and k_2. However, if we restrict[4] k_2 to $-1 < k_2 \leq 1$ we can use Sawa (1972) and Carter (1981) to obtain useful asymptotic expansions of the bias, MSE and risk, in powers of θ^{-1}, where

$$(3.1) \qquad \theta = \frac{(\beta-\beta_o)'X'X(\beta-\beta_o)}{2\sigma^2} = \frac{(\alpha-\alpha_o)'\Lambda(\alpha-\alpha_o)}{2\sigma^2}$$

In common situations the hypothesis error $\beta-\beta_o$, or $\alpha-\alpha_o$, will be non-zero so that θ will be positive. Then, for fixed values of $\beta-\beta_o$ and $\alpha-\alpha_o$, making σ small will make θ large. When this expansion is applied to the bias we have:

$$(3.2) \qquad E(\hat{\beta}-\beta) = \frac{-k_1}{2}\left[\frac{n}{\theta} + \frac{n(n+2)k_2 - nT}{2\theta^2}\right](\beta-\beta_o) + O(\theta^{-3})$$

Both the expansion and the exact bias (equation (A.15) in Appendix A) tell us that $\hat{\beta}$ will become unbiased if $k_1 \to 0$, so that $\hat{\beta} \to b$, or if $\beta \to \beta_o$. In addition the expansion shows us that, to order θ^{-2},

$$(3.3) \qquad \frac{\partial E(\hat{\beta}-\beta)}{\partial k_2} = -\frac{k_1 n(n+2)(\beta-\beta_o)}{4\theta^2}$$

Since $k_1 \geq 0$, this means that k_2 values at one end of its range, say $k_2 = -.99$, if $\beta > \beta_o$ or $k_2 = 1.0$ if $\beta < \beta_o$, will be the most beneficial with respect to bias.

Of course the aim of shrinkage estimation is to minimize risk rather than bias. It is convenient for our purposes to consider the weighted risk $E(\hat{\beta}-\beta)'W(\hat{\beta}-\beta)$, where W is a known K • K matrix of weights with rank $J \leq K$. Equation (2.5) is a special case of this weighted risk obtained by setting $W = I_K$. Using Ullah and Ullah (1978) and Carter (1981) we can write

$$(3.4) \qquad E(\hat{\beta}-\beta)'W(\hat{\beta}-\beta) = \sigma^2 tr[(X'X)^{-1}W] + \frac{n(\beta-\beta_o)'W(\beta-\beta_o)}{\theta} \Psi + O(\theta^{-5})$$

where

$$(3.5) \qquad \Psi = \Psi_1 k_1^2 k_2^2 + \Psi_2 k_1^2 k_2 + \Psi_3 k_1 k_2^2 + \Psi_4 k_1^2 + \Psi_5 k_1 k_2 + \Psi_6 k_1$$

and

$$\Psi_1 = \frac{3(n+2)(n+4)(n+6)}{16\theta^3}$$

$$\Psi_2 = \frac{(n+2)(n+4)}{4\theta^2}\left[1 - \frac{(3T+6-\delta)}{2\theta}\right]$$

$$\Psi_3 = \frac{(n+2)(n+4)}{4\theta^3}\left[3 - \frac{\delta}{2}\right]$$

$$\Psi_4 = \frac{(n+2)}{4\theta}\left[1 + \frac{(\delta/2 - T - 2)}{\theta} + \frac{T(3T+6-2\delta)}{4\theta^2}\right]$$

$$\Psi_5 = \frac{(n+2)}{4\theta^2}\left[4 - \delta + \frac{T(\delta-6)}{\theta}\right]$$

$$\Psi_6 = \frac{(4-2\delta)}{4\theta} + \frac{T}{4\theta^2}\left[\delta - 4 + \frac{(6-\delta)(T-2)}{2\theta}\right]$$

with

$$\delta = \frac{(\beta-\beta_o)'X'X(\beta-\beta_o)}{(\beta-\beta_o)'W(\beta-\beta_o)} tr[(X'X)^{-1}W]$$

$$= \frac{(\alpha-\alpha_o)'\Lambda(\alpha-\alpha_o)}{(\alpha-\alpha_o)'C'WC(\alpha-\alpha_o)} tr(\Lambda^{-1}W)$$

The asymptotic expansions (3.2) and (3.4) become more accurate approximations to the exact expressions as θ grows large. The implication

of a large θ value is more easily understood if we subtract $X\beta_o$ from both sides of the model (2.1) to obtain

$$(3.6) \qquad y - X\beta_o = X(\beta-\beta_o) + \varepsilon$$

Now define a population goodness of fit, analogous to the familiar R^2, for the amended model.

$$(3.7) \qquad \rho^2 = \frac{[E(y-X\beta_o)]'[E(y-X\beta_o)]}{E[(y-X\beta_o)'(y-X\beta_o)]}$$

$$= \frac{(\beta-\beta_o)'X'X(\beta-\beta_o)}{(\beta-\beta_o)'X'X(\beta-\beta_o)+T\sigma^2} = \frac{\theta}{\theta+\frac{T}{2}}.$$

ρ^2 measures the goodness of fit of the model net of the hypothesis. Of course, if $\beta_o = 0$ ρ^2 measures the goodness of fit in the original model. Given T, the fit becomes very tight as θ becomes large. For example if T is 40 a θ value of 42 implies a ρ^2 of .68 while a θ value of 1980 implies a ρ^2 of .99. With $(\beta-\beta_o)$, and hence δ, fixed ρ^2 and θ grow large through σ growing small. Such changes make the asymptotic expansions (3.3) and (3.4) very accurate approximations to the exact expressions. They also reduce the bias of $\hat{\beta}$ and bring its weighted risk closer to that of b so that the reduction in risk obtained by using KK instead of LS becomes very small. The cases of most interest here are those for which θ is small enough to give a worthwhile reduction in weighted risk from using KK yet large enough to make the asymptotic expansions fairly accurate.

The asymptotic expansion (3.4) involves powers and products of the ratios $\frac{n}{2\theta}$ and $\frac{T}{2\theta}$. Therefore, if $\theta > \frac{T}{2}$ the terms of order θ^{-3} and θ^{-4} will be small compared to the term of order θ^{-2} and the weighted risk of $\hat{\beta}$ can be closely approximated by

$$(3.8) \qquad E(\hat{\beta}-\beta)'W(\hat{\beta}-\beta) = \sigma^2 tr[(X'X)^{-1}W] + \frac{n(\beta-\beta_o)'W(\beta-\beta_o)}{4\theta^2}[(n+2)k_1^2 - 2(\delta-2)k_1]$$

$$+ O(\theta^{-3})$$

Note that $E(\hat{\beta}-\beta)'W(\hat{\beta}-\beta) < E(b-\beta)'W(b-\beta)$ only if $\delta > 2$ and

$$(3.9) \qquad 0 < k_1 < \frac{2(\delta-2)}{(n+2)} .$$

The minimum value of (3.8), with respect to k_1, occurs when $k_1 = \frac{(\delta-2)}{(n+2)}$. We want to ensure that θ is large enough that the minimum value of the approximate risk $(\hat{\beta}) > 0$. Note that

$$(3.10) \qquad \sigma^2 tr(X'X)^{-1}W = \frac{(\beta-\beta_o)'W(\beta-\beta_o)\,\delta}{2\theta}$$

Then in order for the minimum of (3.8) to be positive we need

$$(3.11) \qquad \theta > \frac{n(\delta-2)^2}{2\delta(n+2)}$$

Since $\frac{n(\delta-2)^2}{2(n+2)\delta}$ can exceed T when δ is large relative to n, $\theta > T$ is not sufficient for (3.11) to hold. Note that $\theta > T$ implies $\rho^2 > 2/3$.

4. Choosing k_1 and k_2: $W = I$

The values of k_1 and k_2 which minimize the weighted risk (3.4) depend upon the values of θ and of δ, which depends, in turn, upon the values chosen for the matrix W and the vectors β_o or α_o. We begin by setting $W = I$ so that (3.4) becomes

$$(4.1) \qquad E(\hat{\beta}-\beta)'(\hat{\beta}-\beta) = \sigma^2 tr(X'X)^{-1} + n\,\frac{(\beta-\beta_o)'(\beta-\beta_o)}{2\theta}\,\Psi + O(\theta^{-5})$$

with

$$(4.2) \qquad \delta = \frac{(\beta-\beta_o)'X'X(\beta-\beta_o)}{(\beta-\beta_o)'(\beta-\beta_o)}\,tr(X'X)^{-1} = \frac{(\alpha-\alpha_o)'\Lambda(\alpha-\alpha_o)}{(\alpha-\alpha_o)'(\alpha-\alpha_o)}\,tr\,\Lambda^{-1}$$

This is the W matrix used by James and Stein (1961) and Ullah and Ullah (1978). It is the appropriate weight matrix if one is concerned with minimizing the sum of the mean-squared-errors (MSEs) of the individual coefficients.

Assume, temporarily, that δ and θ are also known. Then we could find the minimum-risk-member of the double k-class by choosing the values

of k_1 and k_2 which minimize (4.1) subject to the constraints $-1.0 < k_2 \leq 1.0$ and $k_1 \geq 0$. In practise the second constraint is never binding and the first is binding only when θ is very large in which case k_2, can be set close to -1.0, at -.99 say. We will refer to the estimator which uses the optimal feasible k_1, k_2 pair as OPOP. This estimator is a useful benchmark against which to compare other estimators.

Of course, δ and θ are unknown in practise so OPOP is not a feasible estimator. However several operational rules for selecting k_1 and k_2 are available, based on various a priori specifications with respect to δ and θ. We are interested in establishing the usefulness of these rules as compared to the optimum k_1 and k_2 used in OPOP. One implicit prior specification that underlies most available feasible rules is that $\theta \geq T$, so that (3.8) is a close approximation, and that (3.11) holds so that risk $(\hat{\beta}) > 0$ at its minimum. Equation (3.8) does not contain k_2 and it has traditionally been set equal to one. From (3.8) risk $(\hat{\alpha})$ < risk (a) for k_1 in the range $0 < k_1 < \frac{2(\delta-2)}{n+2}$ and risk $(\hat{\alpha})$ is minimized when $k_1 = \frac{\delta-2}{n+2}$. We will label the estimator which uses this k_1 value DD01. Although this estimator is not operational, because δ is unknown in practise, it is another useful benchmark against which to compare estimators that employ some a priori specification of δ.

Operational estimators based on minimizing (3.8) must employ some a priori specification of the unknown parameter δ. James and Stein suggest replacing δ with its minimum possible value $d_1 = \lambda_1 \text{tr } \Lambda^{-1}$ to obtain $k_1 = (d_1 - 2)/(n+2)$ and $k_2 = 1.0$; we label this estimator JS01. This specification for δ is equivalent to assuming that $(\alpha-\alpha_o)' = (\alpha_1-\alpha_{01},0,\ldots,0)$. An attractive property of this choice of k_1 is that it satisfies (3.9). However, it is always smaller than the optimum value $(\delta-2)/(n+2)$. This suggests finding a value $\bar{\delta}$, for δ such that $(\bar{\delta}-2)/(n+2)$ is less than the

upper bound $2(\delta-2)/(n+2)$. A $\bar{\delta}$ which does this is $2d_1 - 2$ which leads to $k_1 = 2(d_1 - 2)/(n+2)$. We label this estimator, employing $k_2 = 1.0$, DS01.

The $(\alpha-\alpha_o)$ specification discussed above is rather drastic. One alternative is to behave as though $\Lambda = I$ (Sclove (1968) analyzed the orthogonal regressor's case) which implies that $\delta = K$ so that, to order θ^{-2}, risk is minimized by setting k_1 to $\frac{K-2}{n+2}$. We label this case, with $k_2 = 1.0$, DK01. Another alternative is to adopt a more reasonable prior specification[5] for $(\alpha-\alpha_o)$. If we specify that the elements of the vector $(\alpha-\alpha_o)$ are all the same this leads to $\delta = \frac{1}{K}\sum_{i=1}^{K}\lambda_i \mathrm{tr}\Lambda^{-1} = \bar{\lambda}\mathrm{tr}\Lambda^{-1} = \bar{d}$. Then we would set $k_1 = \frac{\bar{d}-2}{n+2}$ and $k_2 = 1.0$. We refer to this estimator as DB01.

Of course the risks of DK01 and DB01 will be less than the risk of LS only if their (positive) k_1 values lie below $2(\delta-2)/(n+2)$. This, in turn, is true only if $\delta > (K+2)/2$, for DK01, or $\delta > (\bar{d}+2)/2$ for DB01. These conditions indicate that shrinkage estimation will be most beneficial in the context of models with large δ values.

One major difficulty with estimators which set $k_2 = 1.0$, like JS01 or DS01, is that the signs of the element of $(\hat{\beta} - \beta_o)$ can turn out to be opposite to those of $(b-\beta_o)$. That is, the shrinkage factor $(1-k_1 c)$ in (2.11) and (2.12) can turn out to be negative. One solution to this problem is the positive part estimator of Baranchik (1964) and Stein (1966) which is equal to $\hat{\beta}$ for all samples for which $(1-k_1 c) \geq 0$ and is equal to b for any samples for which $(1-k_1 c) < 0$. Although this estimator is known to dominate the Stein-rule estimator, expressions for its exact or approximate risk are not available.

Another solution to the sign change problem is to vary k_2, as well as k_1, so as to ensure $(1-k_1 c) \geq 0$. To see what values of k_2 will bring this about note that

$$(4.3) \qquad k_1 c = \frac{k_1}{\frac{(y-X\beta_o)'(y-X\beta_o)}{(y-X\beta_o)'M(y-X\beta_o)} - k_2} \leq \frac{k_1}{1-k_2},$$

since

$$0 \leq (y-X\beta_o)'M(y-X\beta_o) < (y-X\beta_o)'(y-X\beta_o).$$

Then a sufficient condition for $k_1 c \leq 1$, and therefore for $1-k_1 c \geq 0$, is $k_1/(1-k_2) \leq 1$ or $k_1 + k_2 \leq 1$. Given knowledge of δ and θ, an optimum positive shrinkage estimator is obtained by finding the k_1 and k_2 values which minimize (4.1) subject to $k_1 \geq 0$, $-1 < k_2 \leq 1$ and $k_1 + k_2 \leq 1$. We label this estimator OPPS. As we will see below, in some cases neither the second nor the third constraint is binding so that OPPS is identical to OPOP.

If θ is unknown but assumed to be large enough for (3.8) to be useful, then a positive shrinkage estimator will result from setting k_1 by making an a priori specification, as above, which leads to a value for δ and then setting $k_2 = 1-k_1$. This modification to k_2 transforms DD01 to DDPS, JS01 to JSPS, DS01 to DSPS, DK01 to DKPS and DB01 to DKPS. The k_1 values employed by DKPS or DBPS could, for some models, exceed 2.0 in which case $(1-k_1) < -1.0$. However, for the asymptotic expansions (3.4) and (3.8) to be valid, we must restrict k_2 to lie in the range $-1 < k_2 \leq 1$. (Carter (1981).) Therefore, in these cases one should set $k_1 = 1.99$ and $k_2 = -.99$.

One further positive shrinkage estimator is to be found in the development leading to equation (2.7) (see Ullah and Ullah (1978) equation (2.14)). This is to set $k_1 = \frac{1}{n}$ and $k_2 = \frac{n-1}{n}$. We label this DFPS. An attractive property of this choice of k_1 is that it satisfies (3.9) so long as $\delta > 2 + (1/2 + 1/n)$. This condition is not much stricter than the

condition $\delta > 2$ which is required for any k_1 to give smaller risk than LS, to order θ^{-2}. Furthermore the k_1 value of 1/n exceeds that used by JS01 so long as $d_1 < 3 + 2/n$.

In order to numerically assess the effectiveness of these various schemes, both operational and nonoperational, for fixing k_1 and k_2 we can compute the ratio, $\mathfrak{R}$ of the risk of KK to that of LS for given values of T, K, δ and θ. Note that, given W, T and K, the detailed structure of X, β and σ^2 influence the risk of $\hat{\beta}$ only through δ and θ and are, therefore, not of direct interest. Of course if δ is changed with θ fixed there must be a compensating change in σ^2. Table 1 shows values of the relative bias (rel. bias) to order θ^{-2} and the risk ratio, $\mathfrak{R}$ to order θ^{-4} for several interesting cases.

Model 1 has a θ value large enough to make (3.4) and (3.8) good approximations yet not so large as to make KK indistinguishable from LS. The value of δ is higher than both $\bar{d}$ and K so all the operational KK estimators will have smaller risk than LS. Setting k_1 and k_2 to minimize (3.4) (OPOP) produced a rather spectacular reduction in risk, vis-à-vis LS, at the cost of a sizeable increase in relative bias. When the minimization was attempted under the constraint sufficient for positive shrinkage (OPPS) the boundary values of k_1 and k_2 resulted and the reduction in risk was slightly less. When (3.8) was minimized, rather than (3.4), with $k_2 = 1.0$ (DD01) the k_1 value obtained was not much different from that used by OPOP. However, the k_2 value of 1.0 was much different to that used by OPOP so the $\mathfrak{R}$ value was somewhat higher for DD01 than for OPOP. Since the k_1 value used by DD01 exceeded 2.0, the boundary k_1 and k_2 values were used for DDPS, making it the same as OPPS.

Since the operational estimators are all based on attempts to minimize (3.8), rather than (3.4), they should be compared to DD01 (or DDPS) rather than to OPOP. In this light DB01 and DBPS performed very well. Indeed, the $\mathfrak{R}$ value of DB01 was less than that of the non-operational benchmark DD01! The other operational estimators were much less impressive. In no case did DS01, DSPS, JS01 or JSPS produce a risk reduction greater than 1%.

Model 2 is different from Model 1 in only one respect; it has a much lower δ value. The unrestricted optimizing k_1 and k_2 values are both small enough that the sufficient condition for positive shrinkage holds so OPPS was not computed. One would expect that very low k_1 and k_2 values would make KK nearly the same as LS and this is confirmed by the low relative bias and high $\mathfrak{R}$ value for OPOP. Although the other benchmark estimators, DD01 and DDPS, used higher k_2 values, their k_1 values were small and their $\mathfrak{R}$ values were close to one.

Of course, the small value of δ was chosen to highlight the effect this specification can have on the operational estimators. The effect was most pronounced on DB01 and DBPS, which had performed so well when δ was high, but which now have risks more than twice that of LS! Similarly, the risks of DK01 and DKPS also exceed those of LS. Unfortunately, it seems impossible to know in practise whether δ is large or small and, hence, whether or not any of these four operation estimators is preferable to LS. The remaining five operational estimators continued to dominate LS but, like OPOP, they gave a reduction in risk of less than 1%.

Models 3 and 4 differ from Models 1 and 2 in that their θ value is higher leading to a higher value of ρ^2, a value typical in applied econometrics. At this higher θ value the k_1 values used by DD01 will be

closer to those used by OPOP since equation (3.8) is nearly the same as (3.4). However, such a high θ also makes (3.4) and (3.8) only slightly different from risk (LS) so that the considerable risk reductions obtained by using OPOP in Model 1 are no longer available. Of course, the benefits from using one of the operational estimators are smaller still. Note that when δ is small (Model 4) DB01, DBPS, DK01 and DKPS will still have risks greater than LS. Indeed, for this model even OPOP has a risk only .1% less than LS.

Results very similar to those displayed in Table 1 have been obtained for different values of K and T. However, since they lead to no different conclusions they are omitted.

5. The MSE of Individual Coefficients

A researcher who was primarily interested in the MSE of the i^{th} coefficient would choose W to be a matrix with zeros in every position except for the i^{th} main diagonal element which would be set to one. This is the W value which implicitly underlies much of the discussion in Aigner and Judge (1977). For this W (3.4) becomes

$$(5.1) \qquad E(\hat{\beta}_i - \beta_i)^2 = \sigma^2 (X'X)^{ii} + \frac{n(\beta_i - \beta_{oi})^2}{\theta} \Psi + O(\theta^{-5})$$

$$= \frac{\sigma^2}{\lambda_i} + \frac{n(\alpha_i - \alpha_{oi})^2}{\theta} \Psi + O(\theta^{-5})$$

with δ replaced by

$$(5.2) \qquad \delta_i = \frac{(\beta-\beta_o)'X'X(\beta-\beta_o)}{(\beta_i - \beta_{oi})^2} (X'X)^{ii} = \frac{(\alpha-\alpha_o)'\Lambda(\alpha-\alpha_o)}{\lambda_i(\alpha_i-\alpha_{oi})^2}$$

where $(X'X)^{ii}$ is the i^{th} main diagonal element of $(X'X)^{-1}$

Similarly (3.8) becomes

Table 1

Relative Bias and Risk Ratios for Several Models and Estimators: W = I

Model 1

K = 10, T = 36, θ = 72

ρ^2 = .80, δ = 72.8

Estimator	k_1	k_2	rel. bias	R
OPOP	2.48	-.815	-.265	.669
OPPS	1.99	-.99	-.200	.683
DD01	2.53	1.0	-.431	.762
DDPS	1.99	-.99	-.200	.683
DB01	1.09	1.0	-.186	.749
DBPS	1.09	-.09	-.144	.766
DK01	.286	1.0	-.0487	.913
DKPS	.286	.714	-.0459	.916
DFPS	.0385	.9615	-.00651	.987
DS01	.0294	1.0	-.00501	.990
DSPS	.0294	.9706	-.00498	.990
JS01	.0147	1.0	-.00251	.995
JSPS	.0147	.9853	-.00250	.995

Model 2

K = 10, T = 36, θ = 72

ρ^2 = .80, δ = 3.23

k_1	k_2	rel. bias	R
.0656	.0646	-.00903	.995
-	-	-	-
.0438	1.0	-.00747	.997
.0438	.9562	-.00740	.996
1.09	1.0	-.186	2.56
1.09	-.09	-.144	2.11
.286	1.0	-.0487	1.08
.286	.714	-.0459	1.06
.0385	.9615	-.00651	.997
.0294	1.0	-.00501	.997
.0294	.9706	-.00498	.997
.0147	1.0	-.00251	.998
.0147	.9853	-.00250	.998

Model 3

K = 10, T = 36, θ = 342

ρ^2 = .95, δ = 72.8

Estimator	k_1	k_2	rel. bias	R
OPOP	2.53	-.99	-.0872	.914
OPPS	1.99	-.99	-.0686	.918
DD01	2.53	1.0	-.0950	.916
DDPS	1.99	-.99	-.0686	.918
DB01	1.09	1.0	-.0410	.939
DBPS	1.09	-.09	-.0392	.941
DK01	.286	1.0	-.0107	.980
DKPS	.286	.714	-.0106	.981
DFPS	.0385	.9615	-.00144	.997
DS01	.0294	1.0	-.00110	.998
DSPS	.0294	.9706	-.00110	.998
JS01	.0147	1.0	-.000552	.999
JSPS	.0147	.9853	-.000552	.999

Model 4

K = 10, T = 36, θ = 342

ρ^2 = .95, δ = 3.23

k_1	k_2	rel. bias	R
.0545	-.99	-.00188	.999
-	-	-	-
.0438	1.0	-.00165	.999
.0438	.9562	-.00164	.999
1.09	1.0	-.0410	1.35
1.09	-.09	-.0392	1.32
.286	1.0	-.0107	1.02
.286	.714	-.0106	1.02
.0385	.9615	-.0144	.999
.0294	1.0	-.00110	.999
.0294	.9706	-.00110	.999
.0147	1.0	-.000552	.9996
.0147	.9853	-.000552	.9996

$$(5.3)\qquad E(\hat{\beta}_i-\beta_i)^2=\sigma^2(X'X)^{ii}+\frac{n(\beta_i-\beta_{oi})^2}{4\theta^2}\left[(n+2)k_1^2-2(\delta_i-2)k_1\right]+O(\theta^{-3})$$

$$=\frac{\sigma^2}{\lambda_i}+\frac{n(\alpha_i-\alpha_{oi})^2}{4\theta^2}\left[(n+2)k_1^2-2(\delta_i-2)k_1\right]+O(\theta^{-3})$$

from which we can see that the risk of $\hat{\beta}_i$, to $O(\theta^{-3})$, will be less than that of b_i only if $0 < k_1 < 2(\delta_i-2)/(n+2)$. Since one k_1 value will be used in estimating the value of every β_i, $i=1,\ldots,K$, we must choose it to be smaller than the minimum value of $2(\delta_i-2)/(n+2)$ over all i, if we want all $\hat{\beta}_i$ to dominate all b_i. In the spirit of James and Stein, we might seek a values $d_{1m} \leq \delta_m$, where δ_m is the smallest of the δ_i values, and set $k_1=(d_{1m}-2)/(n+2)$. However, we can see from the last term on the right of (5.2), that

$$(5.4)\qquad \delta_i = 1+\frac{\sum_{j\neq i}^{K}\lambda_j(\alpha_j-\alpha_{oj})^2}{\lambda_i(\alpha_i-\alpha_{oi})^2}$$

so that $d_{1m}=1$ which would lead to a negative value for k_1. That is, there is no positive k_1 value which is guaranteed to be less than $2(\delta_i-2)/(n+2)$ for all i. In particular, the k_1 values discussed in the previous section, chosen to minimize risk with $W=I$, will give ratios of $MSE(\hat{\alpha}_i)/MSE(a_i) > 1$ for at least one i.

If the hypothesis about a particular coefficient is nearly correct, $\alpha_i-\alpha_{oi}$ will be small so δ_i will tend to be large and the ratio $MSE(\hat{\alpha}_i)/MSE(a_i)$ small. Of course, this will happen more readily if δ is small which will occur if large hypothesis errors $(\alpha_i-\alpha_{oi})$ are associated with small roots λ_i and vice versa. However, since this can be false for models with widely differing λ_i and $(\alpha_i-\alpha_{oi})$ values, it is easily possible for $MSE(\hat{\alpha}_i)$ to exceed $MSE(a_i)$. This would be very undesirable to a researcher concerned with individual coefficient estimates as well as with overall risk. Given the δ_i values, such a researcher could set k_1 and k_2 so as to minimize the

MSE of the coefficient with the smallest of them. However, this rule can break down if the smallest δ_i is less than 2.0 because then the minimizing k_1 will be negative to order θ^{-2} and perhaps also to order θ^{-4}. In this case all $k_1 > 0$ will lead to $MSE(\hat{\alpha}_i) > MSE(a_i)$. Also, knowledge of all δ_i values is equivalent to knowledge of all β_i values!

Table 2 provides a numerical illustration using Models 1 and 2 of Table 1 for the best of the benchmark estimators and the two best operational estimators. Recall that for Model 2, OPOP and OPPS and the same. The numbers in the body of the table are ratios $MSE(\hat{\alpha}_i)/MSE(a_i)$. The results of Table 2 illustrate the fact that when OPPS achieves spectacular risk reduction over LS it does so at the cost of increasing the MSE for several coefficients. However, when δ is small with respect to θ, reductions in risk from OPPS are modest but its MSE is less than that of LS so long as $\delta_i > 2$. The two operational estimators fare quite well with regard to individual MSEs, although they work poorly when the minimum $\delta_i < 2$. The reason for their comparative success in terms of individual coefficient MSEs is that they employ comparatively small k_1 values.

Whether the ratio $MSE(\hat{\alpha}_i)/MSE(a_i)$ exceeds one for any particular coefficient depends not only upon $(\alpha_i - \alpha_{oi})$ for that coefficient but also upon that same difference for every other coefficient in the model; e.g., compare the ratios for coefficient 10 in Models 1 and 2. Since this pattern is unknown, a priori, one cannot say for which coefficients of a multiple regression $\hat{\alpha}_i$ dominates a_i.

The value of the W matrix used here effectively reduces the number of coefficients being estimated from ten to one. The results of Stein (1961) and Ullah and Ullah (1978) indicate that KK will dominate LS only if $K \geq 3$ so any W matrix of rank less three, will lead to increased risk vis-à-vis LS.

Table 2

MSE Ratios for Individual Coefficients of Models 1 and 2

Model 1

$K = 10,\ T = 36,\ \theta = 72,\ \rho^2 = .80,\ \delta = 72.8$

$\alpha_i - \alpha_{oi}$	δ_i	OPPS	DFPS	DSPS
100.0	765000	.471	.987	.990
81.0	239000	.472	.987	.990
64.0	3016	.477	.987	.990
49.0	613	.497	.987	.990
36.0	138	.583	.987	.990
25.0	46.2	.805	.988	.991
16.0	22.4	1.16	.988	.991
9.0	10.7	1.92	.990	.992
4.0	5.57	3.24	.993	.994
1.0	1.54	10.5	1.007	1.005

Model 2

$K = 10,\ T = 36,\ \theta = 72,\ \rho^2 = .80,\ \delta = 3.23$

$\alpha_i - \alpha_{oi}$	δ_i	OPPS	DFPS	DSPS
19.0	1.37	1.015	1.010	1.007
1.0	248	.981	.987	.990
1.0	158	.981	.987	.990
1.0	102	.981	.987	.990
1.0	55.8	.982	.988	.990
1.0	38.8	.982	.988	.991
1.0	34.9	.982	.988	.991
1.0	28.3	.982	.988	.991
1.0	23.6	.983	.988	.991
1.0	9.94	.986	.990	.992

6. Linear Combinations of Coefficients

The third case we will consider is that in which interest is focussed on a hypothesis like (2.9) in which J, the number of linearly independent rows in R, is at least three. Under this null hypothesis, the quadratic form $\sigma^{-2}(b-\beta)'R'[R(X'X)^{-1}R']^{-1}R(b-\beta)$ has a χ^2 distribution with J degrees of freedom. Given σ, this leads to confidence regions for $R\beta$ whose size depends, in part, on the expectation of this quadratic form. This motivates replacing b by $\hat{\beta}$ and using $R'[R(X'X)^{-1}R']^{-1}R$ as W in (3.4). A unique property of this W is that changes in R, i.e., changes in H_o, change W, and hence δ, even though X and β may be constant.

For a fixed value of R (with $J > 3$) knowledge of δ and θ can be used to find the values of k_1 and k_2 which minimize (3.4) subject to $k_1 > 0$, $-1 < k_2 \leq 1$, which we again label OPOP. Adding the constraint $k_1 + k_2 \leq 1$ ensures that the elements of Rb and $R\hat{\beta}$ have the same sign and we retain the label OPPS for this estimator.

In this case too we can assume that θ is large enough to justify basing our operational choice of k_1 upon (3.8). As before, the quality of our estimator will depend upon the a priori specification of δ which is used in forming k_1. As a first step in specifying δ, note that $tr[(X'X)^{-1}W] = J$ in this case. When $J < K$ the matrix $X'X - R'[R(X'X)^{-1}R']^{-1}R$ is positive semi-definite so that $\delta > J$. This suggests, in the spirit of James and Stein, setting $k_1 = (J-2)/(n+2)$ and $k_2 = 1.0$ or $k_2 = 1-k_1$. We retain the labels JS01 and JSPS, respectively, for these estimators. This value of k_1 satisfies (3.9) but it is always smaller than the optimum so we seek a larger k_1 value guaranteed to be smaller than $2(\delta-2)/(n+2)$. A value which does this is $k_1 = 2(J-2)/(n+2)$ which leads to the two estimators DS01, when $k_2 = 1.0$, and DSPS, when $k_2 = 1-k_1$.

An important special case arises when the hypothesis involves all the elements of β so that $J = K$. In this case $\delta = K$ which, of course, is known exactly! An example of this case arises if all the variables are measured as deviations about their sample averages so that β is composed entirely of slope coefficients. Then one would typically test the hypothesis that the elements of β were jointly zero:[6] i.e., $I\beta = 0$. Even if $J < K$, δ will still equal K if all of the following conditions hold: $X'X = I$, the columns of R can be arranged so that $R = [I_J: \ 0_{K-J}]$ (i.e., a partitioned matrix with one portion a J order identity matrix and the remaining K-J columns all zeros), and all elements of the β vector are the same. Although all these conditions will be met only rarely, they may be approximately true in enough cases to justify setting $k_1 = (K-2)/(n+2)$ as a general rule, leading to the two estimators DK01 and DKPS.

The specification $k_1 = \frac{1}{n}$ is not useful here because it leads to a k_1 value less than $(J-2)/(n+2)$, which is itself typically smaller than the optimum k_1, for all J values greater than $3 + 2/n$.

Table 3 presents numerical results for a number of cases. Here the term "rel. bias" refers to the relative bias of $R\beta$ and, as before, $\mathcal{R}$ shows the weighted risk of $\hat{\beta}$ relative to that of b. The first section of the table is concerned with Model 1, which also appeared in Tables 1 and 2. The ten cases numbered (i) to (x) represent ten combinations of J and δ which could arise from ten different R matrices (i.e., ten different null hypotheses). In many of these cases OPOP is the same as OPPS because its $k_1 + k_2 < 1.0$. When J and δ are both small (Model 1 cases (i) and (ii)) the values of $\mathcal{R}$ are close to one even for OPOP. In case (i) $\delta < (K+2)/2$ so the $\mathcal{R}$ values for DK01 and DKPS exceed one. As δ grows larger, with J and the model

structure fixed, the potential reduction in risk of KK over LS becomes larger; see Model 1 cases (iii) and (iv). Unfortunately, the operational estimators capture, at most, only about a third of this potential gain. DK01 and DKPS are best in this respect, while JS01 and JSPS offer only very small gain. DS01 and DSPS are less effective than DK01 and DKPS when J is very small and δ is very large. However, they have the advantage of always having $\mathcal{R}$ values less than, or equal to, one; see Model 1 case (i). Also, when $J > (K+2)/2$ the $\mathcal{R}$ values of DS01 and DSPS are less than those of DK01 and DKPS, so long as δ is large enough (see Model 1 cases (v), (vi) and (viii)) because the k_1 values for the former two estimators exceed those of the later two and are closer to the optimum value. Of course, if δ is only slightly larger than J the larger k_1 value for DS01 and DSPS are a disadvantage vis-à-vis DK01 and DKPS (see Model 1 case (vii)), although the range of circumstances under which this will happen seems small.

For Model 1 cases (ix) and (x) δ is the same although the J values are different. Therefore, OPOP, DK01 and DKPS are all the same for these two cases. Of course, in case (x) where $J = K$ DS01 and DSPS would never be used and JS01 and JSPS are the same as DK01 and DKPS. Since δ is known in this case almost all the potential reduction in risk exhibited by OPOP is realized by the operational estimator. The lack of knowledge of θ, which prevents OPOP from being operational in these cases, makes very little difference.

The results for Models 5 to 8, which appear in the second part of Table 3 also illustrate these findings. Models 5 and 6 have the same degrees of freedom and goodness of fit as Model 1 and in both cases $J = K$. The $\mathcal{R}$ ratios are smaller when δ is larger and the operational estimators have $\mathcal{R}$ values very nearly as small as those for the benchmark estimators.

Models 7 and 8 were obtained by halving the degrees of freedom of Models 5 and 6 with everything else held constant. This had the effect of reducing the $\mathfrak{R}$ values for all the estimators while leaving unchanged the proximity of operational and benchmark estimators and the relatively small $\mathfrak{R}$ values produced by a large δ.

Model 3 also appeared in Table 1. It has a higher goodness of fit than Model 1 but is the same in all other respects. Results for J and δ values of cases (i), (iv), (ix) and (x) are tabled. The $\mathfrak{R}$ values for these cases, when compared to Model 1, are all much closer to one, even those for DK01 and DKPS which exceed one. In addition the absolute value of the relative bias is much reduced. This same trend is evident from the results for Models 9 and 10 which are Models 7 and 8 with the goodness of fit increased.

Table 3

Relative Biases and Risk Ratios for Linear Combinations of Coefficients

Model 1

$K = 10,\ T = 36,\ \theta = 72,\ \rho^2 = .80$

Estimator	(i) J = 3 δ = 5 k_1	k_2	rel. bias	ℜ	(ii) J = 3 δ = 7 k_1	k_2	rel. bias.	ℜ
OPOP	.138	.155	-.0195	.986	.218	.208	-.0311	.975
DK01	.286	1.0	-.0487	1.016	.286	1.0	-.0487	.984
DKPS	.286	.714	-.0459	1.009	.286	.714	-.0459	.981
DS01	.0714	1.0	-.0122	.990	.0714	1.0	-.0122	.986
DSPS	.0714	.929	-.0120	.990	.0714	.929	-.0120	.986
JS01	.0357	1.0	-.00609	.994	.0357	1.0	-.00609	.992
JSPS	.0357	.964	-.00605	.994	.0357	.964	-.00605	.992
	(iii) J = 3 δ = 13				(iv) J = 3 δ = 60			
OPOP	.447	.248	-.0644	.941	2.05	-.510	-.241	.723
OPPS	-	-	-	-	1.93	-.931	-.198	.728
DK01	.286	1.0	-.0487	.948	.286	1.0	-.0487	.914
DKPS	.286	.714	-.0459	.948	.286	.714	-.0459	.918
DS01	.0714	1.0	-.0122	.981	.0714	1.0	-.0122	.977
DSPS	.0714	.929	-.0120	.981	.0714	.929	-.0120	.977
JS01	.0357	1.0	-.00609	.990	.0357	1.0	-.00609	.988
JSPS	.0357	.964	-.00605	.990	.0357	.964	-.00605	.988
	(v) J = 7 δ = 13				(vi) J = 7 δ = 60			
OPOP	.447	.248	-.0644	.941	2.05	-.510	-.241	.723
OPPS	-	-	-	-	1.93	-.931	-.198	.728
DK01	.286	1.0	-.0487	.948	.286	1.0	-.0487	.914
DKPS	.286	.714	-.0459	.948	.286	.714	-.0459	.918
DS01	.357	1.0	-.0609	.944	.357	1.0	-.0609	.895
DSPS	.357	.643	-.0564	.943	.357	.643	-.0564	.901
JS01	.179	1.0	-.0305	.960	.179	1.0	-.0305	.944
JSPS	.179	.821	-.0293	.961	.179	.821	-.0293	.946

Table 3 (cont'd.)

Model 1 (cont'd.)

Estimator	(vii) J = 9 δ = 11 k_1	k_2	rel. bias	ℜ	(viii) J = 9 δ = 15 k_1	k_2	rel. bias	ℜ
OPOP	.372	.247	-.0536	.953	.521	.243	-.0749	.930
DK01	.286	1.0	-.0487	.955	.286	1.0	-.0487	.942
DKPS	.286	.714	-.0459	.955	.286	.714	-.0459	.943
DS01	.500	1.0	-.0853	.969	.500	1.0	-.0853	.934
DSPS	.500	.500	-.0765	.960	.500	.500	-.0765	.931
JS01	.250	1.0	-.0426	.957	.250	1.0	-.0426	.946
JSPS	.250	.750	-.0404	.957	.250	.750	-.0404	.947
	(ix) J = 5 δ = 10				(x) J = 10 δ = 10			
OPOP	.334	.242	-.0481	.958	.334	.242	-.0481	.958
DK01	.286	1.0	-.0487	.960	.286	1.0	-.0487	.960
DKPS	.286	.714	-.0459	.959	.286	.714	-.0459	.959
DS01	.214	1.0	-.0365	.963	-	-	-	-
DSPS	.214	.786	-.0349	.963	-	-	-	-
JS01	.107	1.0	-.0183	.976	-	-	-	-
JSPS	.107	.893	-.0179	.976	-	-	-	-
	Model 5 K = 6, T = 32, θ = 64, ρ^2 = .80 J = 6 δ = 6				Model 6 K = 24, T = 50, θ = 100, ρ^2 = .80 J = 24 δ = 24			
OPOP	.178	.165	-.0285	.978	.857	.325	-.0887	.914
OPPS	-	-	-	-	.864	.136	-.0864	.914
DK01	.143	1.0	-.0281	.981	.786	1.0	-.0909	.915
DKPS	.143	.857	-.0272	.980	.786	.214	-.0797	.914
	Model 7 K = 6, T = 19, θ = 38, ρ^2 = .80 J = 6 δ = 6				Model 8 K = 24, T = 37, θ = 74, ρ^2 = .80 J = 24 δ = 24			
OPOP	.332	.138	-.0442	.966	1.59	.155	-.107	.894
OPPS	-	-	-	-	1.56	-.559	-.0950	.896
DK01	.267	1.0	-.0432	.970	1.47	1.0	-.110	.896
DKPS	.267	.733	-.0408	.969	1.47	-.467	-.0905	.896

Table 3 (cont'd.)

Model 3

$K = 10,\ T = 36,\ \theta = 342,\ \rho^2 = .95$

Estimator	(i) $J = 3\ \ \delta = 5$ k_1	k_2	rel. bias	R	(iv) $J = 3\ \ \delta = 60$ k_1	k_2	rel. bias	R
OPOP	.126	-.99	-.00426	.997	2.11	-.99	-.0727	.929
OPPS	-	-	-	-	1.99	-.99	-.0686	.929
DK01	.286	1.0	-.0107	1.004	.286	1.0	-.0107	.981
DKPS	.286	.714	-.0106	1.004	.286	.714	-.0106	.981
DS01	.0714	1.0	-.00268	.998	.0714	1.0	-.00268	.995
DSPS	.0714	.929	-.00268	.998	.0714	.929	-.00268	.995
JS01	.0357	1.0	-.00134	.999	.0357	1.0	-.00134	.997
JSPS	.0357	.964	-.00134	.999	.0357	.964	-.00134	.997
	(ix) $J = 5\ \ \delta = 10$				(x) $J = 10\ \ \delta = 10$			
OPOP	.317	-.99	-.0109	.991	.317	-.99	-.0109	.991
DK01	.286	1.0	-.0107	.991	.286	1.0	-.0107	.991
DKPS	.286	.714	-.0106	.991	.286	.714	-.0106	.991
DS01	.214	1.0	-.00805	.992	-	-	-	-
DSPS	.214	.786	-.00798	.992	-	-	-	-
JS01	.107	1.0	-.00403	.995	-	-	-	-
JSPS	.107	.893	-.00401	.995	-	-	-	-

Model 9

$K = 6,\ T = 19,\ \theta = 180.5,\ \rho^2 = .95$

$J = 6\ \ \delta = 6$

Estimator	k_1	k_2	rel. bias	R
OPOP	.303	-.99	-.00992	.993
DK01	.267	1.0	-.00950	.994
DKPS	.267	.733	-.00939	.994

Model 10

$K = 24,\ T = 37,\ \theta = 351.5,\ \rho^2 = .95$

$J = 24\ \ \delta = 24$

Estimator	k_1	k_2	rel. bias	R
OPOP	1.55	-.99	-.0265	.975
DK01	1.47	1.0	-.0263	.976
DKPS	1.47	-.47	-.0254	.976

7. Forecasting

In many cases the thing of most interest is $X_f\beta$, where X_f is a matrix of known future values of the explanatory variables. If the number of rows of X_f is $3 \leq J \leq K$ and one wishes to form confidence intervals for $X_f\beta$ then X_f is just an example of R in Section 6 and the results of that section carry over directly. If one desires to make joint forecasts over at most two future periods then LS should be employed. Joint forecast intervals over more than K independent future periods are difficult for both LS and KK because in this case $X_f(X'X)^{-1}X_f'$ is singular.

A situation close to this arises when one attempts to estimate the mean of y conditional on X, $X\beta$. The LS estimator now is $Xb = Za$ which has a singular covariance matrix $\sigma^2 X(X'X)^{-1}X' = \sigma^2 Z\Lambda^{-1}Z'$ and a risk of $\sigma^2 K$. So long as $3 \leq K < T$, the estimator $X\hat{\beta}$ has a smaller risk which is (3.4), or (3.8), with $W = X'X$. This value for W makes $\delta = K$ and so it is analogous to the case of $J = K$ in Section 6. As in that case, almost all the potential gain from using the optimum k values, OPOP, will be obtained by the operational estimators DK01 and DKPS. The portions of Table 3 in which $J = K$ can be consulted for illustrations.

8. Conclusions

In the context of the classical, normal, linear regression model we have considered the double k-class estimators of: individual coefficients, the whole coefficient vector and linear combinations of the elements of the coefficient vector, which includes the expectation of y conditional on future and current X values.

KK should not be employed if interest is focussed primarily on individual coefficients because it is impossible to know, a priori, whether the MSE of the KK estimator of a particular coefficient is greater or less than its LS estimator.

On the other hand, if one desires to minimize the sum of the individual MSE's, KK can be usefully employed. Knowledge of the value of the population parameter δ could be employed to give an estimator with much smaller risk than LS if δ is large and the population goodness of fit is moderately tight (about .80). In practise, setting $k_1 = \frac{1}{n}$ and $k_2 = 1 - k_1$ will give a small reduction in risk over LS in almost all parts of the parameter space, without changing the sign of the LS estimates. Over a fairly wide portion of the parameter space ($\delta > (K+2)/2$) more substantial gains can be obtained by changing k_1 to $(K-2)/(n+2)$, although this will lead to risks greater than LS if δ is fairly small. Other values for k_1 can be found, including that suggested by James and Stein, which lead to KK dominating LS everywhere but the gains obtained are miniscule.

KK seems most useful in estimating three or more independent linear combinations of the regression coefficients in the contexts of joint tests of hypothesis about the coefficients or joint conditional forecasts of y. So long as J, the number of linear combinations, is less than K, the practical procedure is to set $k_1 = 2(J-2)/(n+2)$ and $k_2 = 1-k_1$ for positive shrinkage. The KK estimator obtained will always have smaller risk than LS with the gain being up to 10% (noticeably larger than in the

previous case) when δ is large relative to J and the goodness-of-fit is moderately tight. If, as may often be the case, the number of linear combinations (or number of forecast periods) considered is equal to the number of coefficients $\delta = K$ so that k_1 should be changed to $(K-2)/(n+2)$. This value of k_1 should also be used in estimating the expectation of y conditional on the sample X. The reduction in risk produced here will be slightly less than the case in which δ is large relative to J but the operational estimator achieves nearly all the risk reduction which is theoretically possible.

If the population goodness-of-fit increases to .95 or more the amount of risk reduction which is possible using KK is very small even if values of the two population parameters δ and θ are known. In such cases the operational estimators lead to even smaller risk reductions. There is also very little risk reduction obtainable when δ is small or when the degrees of freedom are large. In summary, KK is a technique which is most useful when the degrees of freedom are small, the model is believed to fit loosely and one is interested in K independent linear combinations of the K regression coefficients.

Appendix A

Judge, Griffiths, Hillard and Lee (1980, p. 68) have presented a version of the Stein-rule estimator which shrinks b towards β_o. It uses the likelihood ratio test statistic

$$(A.1) \qquad u = \frac{(T-K)(b-\beta_o)'X'X(b-\beta_o)}{Ke'e}$$

in

$$(A.2) \qquad \tilde{\beta} = (1 - \frac{a_1}{u})(b-\beta_o) + \beta_o$$

$$= [1 - \frac{k_1 e'e}{(b-\beta_o)'X'X(b-\beta_o)}](b-\beta_o) + \beta_o$$

$$= [1 - \frac{k_1 e'e}{(y-X\beta_o)'(y-X\beta_o) - e'e}](b-\beta_o) + \beta_o$$

where a_1 and k_1 are arbitrary scalars.

This naturally suggests the following double k-class estimator which shrinks b towards β_o rather than towards zero

$$(A.3) \qquad \hat{\beta} = [1 - \frac{k_1 e'e}{(y-X\beta_o)'(y-X\beta_o) - k_2 e'e}](b-\beta_o) + \beta_o$$

$$= (1 - k_1 c)(b-\beta_o) + \beta_o,$$

or

$$(A.4) \qquad \hat{\alpha} = (1-k_1 c)(a-\alpha_o) + \alpha_o$$

where $\alpha_o = C'\beta_o$,

$$(A.5) \qquad c = \frac{e'e}{(y-X\beta_o)'(y-X\beta_o) - k_2 e'e}$$

$$= \frac{(y-X\beta_o)'M(y-X\beta_o)}{(y-X\beta_o)'(y-X\beta_o) - k_2(y-X\beta_o)'M(y-X\beta_o)},$$

and $M = I - X(X'X)^{-1}X'$. The special cases (2.7) and (2.8) are obtained from (A.3) and (A.4) by setting $\beta_o = \alpha_o = 0$.

The sampling error of $\hat{\beta}$ is

$$\text{(A.6)} \qquad \hat{\beta} - \beta = (b-\beta) - k_1 c(b-\beta_o).$$

Now define

$$\text{(A.7)} \qquad z = \frac{1}{\sigma} P'(y-X\beta_o) \sim N[\frac{1}{\sigma} P'X(\beta-\beta_o), I]$$

where P is the orthogonal matrix of characteristic vectors of M. Then c can be written as

$$\text{(A.8)} \qquad c = \frac{z'D_1 z}{z'D_2 z}$$

where $D_1 = P'MP = \begin{bmatrix} I_n & 0 \\ 0 & 0 \end{bmatrix}$, $D_2 = I - k_2 D_1$ and $n = T-K$. Then we can follow the proofs of Theorems 1, 2 and 3 in Ullah and Ullah (1978) to obtain the bias, MSE matrix and risk of $\hat{\beta}$.

First we define

$$\text{(A.9)} \qquad g_{u,v} = \int_{-\infty}^{0} 2\exp\{\frac{2\theta t}{1-2t}\}(1-2t)^{\frac{n-T}{2}+v-u}[1-2(1-k_2)t]^{\frac{n}{2}-v} dt$$

$$= \exp(-\theta) \sum_{i=0}^{\infty} \frac{\theta^i}{i!(\frac{T}{2}+u+i-1)} {}_2F_1(1, \frac{n}{2}+v; \frac{T}{2}+u+i; k_2)$$

where

$$\text{(A.10)} \qquad \theta = \frac{(\beta-\beta_o)'X'X(\beta-\beta_o)}{2\sigma^2} = \frac{(\alpha-\alpha_o)'\Lambda(\alpha-\alpha_o)}{2\sigma^2}$$

and ${}_2F_1(1,p;q;w)$ is a hypergeometric function which can be written as

$$(A.11) \qquad {}_2F_1(1,p;q;w) = \sum_{j=o}^{\infty} \frac{(p)_j}{(q)_j} w^j$$

with $(p)_j$ and $(q)_j$ ascending factorials: i.e., $(p)_j = 1 \cdot 2 \cdot 3 \cdot \ldots (p+j-1)$. Note that

$$(A.12) \qquad \lim_{\theta \to 0} g_{u,v} = \left(\frac{2}{T+2u-2}\right) {}_2F_1\left(1, \frac{n}{2}+v; \frac{T}{2}+u; k_2\right)$$

so that

$$(A.13) \qquad \lim_{(\beta-\beta_o) \to 0} g_{u,v}(\beta-\beta_o) = 0, \text{ a } K \cdot 1 \text{ vector of zeros,}$$

and

$$(A.14) \qquad \lim_{(\beta-\beta_o) \to 0} g_{u,v}(\beta-\beta_o)(\beta-\beta_o)' = 0, \text{ a } K \cdot K \text{ matrix of zeros}$$

Now, following Ullah and Ullah (1978), Theorem 1

$$(A.15) \qquad E(\hat{\beta}-\beta) = -\frac{nk_1}{2} g_{2,1}(\beta-\beta_o)$$

Given n and $(\beta-\beta_o)$, variations in k_1 and k_2 have the same effect upon the relative bias of each element of $\hat{\beta}$: the relative bias becomes more negative as k_1 and/or k_2 increase (with $k_2 \leq 1$) (Ullah and Ullah (1978), Theorem 1, Corollary 2(b)). Given n and k_1, the size and direction of the bias of individual elements of $\hat{\beta}$ depend upon the size and direction of the difference between that coefficient's hypothesized value and its true value: individual elements of $\hat{\beta}$ will be unbiased if their hypothesized values are true. As the elements of β_o all get closer to β, equation (A.13) shows that $\hat{\beta}$ becomes unbiased.

Following Ullah and Ullah (1978), Theorem 2, the MSE matrix of $\hat{\beta}$ is

$$(A.16)\quad E(\hat{\beta}-\beta)(\hat{\beta}-\beta)' = \sigma^2(X'X)^{-1} - k_1 n[g_{2,1} + \frac{k_1(n+2)}{4}(g_{3,1} - g_{2,2})]\sigma^2(X'X)^{-1}$$
$$- k_1 n[g_{3,1} - g_{2,1} + \frac{k_1(n+2)}{4}(g_{4,2} - g_{3,2})](\beta-\beta_o)(\beta-\beta_o)'$$

To obtain the risk of $\hat{\beta}$ we take the trace of (A.16) to obtain

$$(A.17)\quad E(\hat{\beta}-\beta)'(\hat{\beta}-\beta) = E(\hat{\alpha}-\alpha)'(\hat{\alpha}-\alpha)$$
$$= tr\Lambda^{-1} - k_1 n[g_{2,1} + \frac{k_1(n+2)}{4}(g_{3,2} - g_{2,2})]\sigma^2 tr\Lambda^{-1}$$
$$- k_1 n[g_{3,1} - g_{2,1} + \frac{k_1(n+2)}{4}(g_{4,2} - g_{3,2})](\alpha-\alpha_o)'(\alpha-\alpha_o)$$

The size of the MSE matrix and of the risk depend not only upon k_1 and k_2 (through the $g_{u,v}$ terms) but also upon the size of the hypothesis error $\beta-\beta_o$, or $\alpha-\alpha_o$. As this error grows smaller, with all else constant we have

$$(A.18)\quad \lim_{(\beta-\beta_o)\to 0} E(\hat{\beta}-\beta)(\hat{\beta}-\beta)' = \sigma^2(X'X)^{-1} - k_1 n[(\frac{2}{T+2})_2F_1(1, \frac{n}{2}+1; \frac{T}{2}+2; k_2)$$
$$+ \frac{k_1(n+2)}{4}\{(\frac{2}{T+4})_2F_1(1, \frac{n}{2}+2; \frac{T}{2}+3; k_2)$$
$$- (\frac{2}{T+2})_2F_1(1, \frac{n}{2}+2; \frac{T}{2}+2; k_2)\}]\sigma^2(X'X)^{-1}$$

and

$$(A.19)\quad \lim_{(\alpha-\alpha_o)\to 0} E(\hat{\alpha}-\alpha)'(\hat{\alpha}-\alpha) = tr[\lim_{(\beta-\beta_o)\to 0} E(\hat{\beta}-\beta)(\hat{\beta}-\beta)'].$$

Footnotes

[1]I am grateful to A. Ullah for stimulating discussion, to B. H. Bentley for expert programming and research assistance and to the Social Science and Humanities Research Council of Canada for financial assistance in the form of a Leave Fellowship and a Research Grant. I have also received helpful comments from C. Beach, D. Hendry, B. McCabe, A. Nakamura, M. Nakamura, G. D. A. Phillips, N. E. Savin, L. J. Slater and J. Wolters. Much of the research on this topic was completed while I was a visitor at the University of Cambridge.

[2]If $k_2 = 1.0$ the bias exists only if $K \geq 2$ and the mean squared error matrix and risk exist only if $K \geq 3$.

[3]Note that the estimators $\tilde{\beta}$ (A.2) or $\hat{\beta}$ (2.12) shrink b towards the fixed, non-random hypothesized point β_o rather than towards a random restricted least squares estimator $b^* = b + (X'X)^{-1} R' [R(X'X)^{-1} R']^{-1}(r-Rb)$. Whenever $J = K$, so that R is non-singular, $b^* = R^{-1}r$ which is non-random. If $J < K$ and the linear restriction is transformed canonically so that the transformed R is non-singular (see Judge and Bock (1978) p. 84) the restricted least squares estimator in the canonical space is a non-random vector of J elements.

[4]The exact risk of $\hat{\beta}$ and $\hat{\alpha}$ exist for $k_2 \leq -1.0$ but not for $k_2 > 1.0$. The asymptotic expansion of the confluent hypergeometric function given by Sawa (1972) is valid for real values of its parameters. A more general expansion which allows for complex parameters is given by L. J. Slater (1960) p. 60.

[5]It could be argued that a priori specifications ought to be introduced using Baye's Theorem. This would lead to posterior distribution for β of the sort discussed by Tiao and Zellner (1964) or Lindley and Smith (1972).

[6]The distribution of test statistics employing $\hat{\beta}$ is discussed in Ullah, Carter and Srivastava (1983).

References

Aigner, D. J. and G. C. Judge (1977), "Application of Pre-Test and Stein Estimators to Economic Data," Econometrica, Vol. 45, pp. 1279-1288.

Baranchik, A. J. (1964), "Multiple Regression and Estimation of the Mean of a Multivariate Normal Distribution," Technical Report 51 (Stanford, Department of Statistics).

Bock, M. E. (1975), "Minimax Estimators of the Mean of a Multivariate Normal Distribution," Annals of Mathematical Statistics, Vol. 3, pp. 209-218.

Carter, R. A. L. (1981), "Improved Stein-Rule Estimator for Regression Problems," Journal of Econometrics, Vol. 17, pp. 113-123, and "Erratum," Journal of Econometrics, Vol. 17, pp. 393-394.

James, W. and C. Stein (1961), "Estimation with Quadratic Loss," Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, pp. 361-379, Berkeley: University of California Press.

Judge, G. G. and W. E. Bock (1976), "A Comparison of Traditional and Stein-Rule Estimators Under Weighted Squared Error Loss," International Economic Review, Vol. 17, pp. 234-240.

Judge, G. G. and M. E. Bock (1978), The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics, Amsterdam: North-Holland Publishing Company.

Judge, G. G., W. E. Griffiths, R. C. Hill, and F. C. Lee (1980), The Theory and Practice of Econometrics, New York: John Wiley and Sons.

Lindley, D. V. and A. F. M. Smith (1972), "Bayes Estimates for the Linear Model," Journal of the Royal Statistical Society, B, Vol. 34, pp. 1-18.

Sawa, T. (1972), "Finite Sample Properties of the k-Class Estimators," Econometrica, Vol. 40, pp. 653-680.

Sclove, S. L. (1968), "Improved Estimators for Coefficients in Linear Regression," Journal of the American Statistical Association, Vol. 63, pp. 597-606.

Slater, L. J. (1960), Confluent Hypergeometric Functions, Cambridge: The University Press.

Stein, C. (1956), "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution," Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, pp. 197-206, Berkeley: University of California Press.

Stein, C. (1966), "An Approach to the Recovery of Inter-Block Information on Balanced Incomplete Block Designs," F. H. David (ed.), Research Papers in Statistics, New York: John Wiley and Sons, pp. 351-366.

Tiao, G. C. and A. Zellner (1964), "On the Bayesian Estimation of Multivariate Regression," Journal of the Royal Statistical Society, B, Vol. 26, pp. 277-285.

Ullah, A. and S. Ullah (1978), "Double k-Class Estimators of Coefficients in Linear Regression," Econometrica, Vol. 46, pp. 705-722, and "Errata".

Ullah, A., R. A. L. Carter and V. K. Srivastava (1983), "Sampling Distribution of Shrinkage Estimators and Their F-Ratios in the Regression Model," Department of Economics, University of Western Ontario, Research Report No. 8324.