ON THE USE OF 'IMPROVED' ESTIMATORS IN ECONOMETRICS

Benny M.S. Lee

A thesis submitted to the
Australian National University
for the degree of
Doctor of Philosophy,
September, 1979.

DECLARATION

Except where otherwise acknowledged
the work described here is my own.

B.M.S. Lee

ACKNOWLEDGEMENT

I would like to thank all the members of the Statistics Department, School of General Studies, for their many helpful comments throughout the course of my study. In particular, I am indebted to my friend and thesis supervisor, Dr P.K. Trivedi, for his patient guidance and for alerting me to the problem of biased estimation. Professor R.D. Terrell is a constant source of help and encouragment and I am especially grateful for his reading and criticising on the first draft of this thesis.

The financial support and the pleasant environment provided by the Australian National University must be acknowledged as one of the most important factors which enable the completion of this study.

Finally, I would like to thank Ms H. Patrikka for the professional service she rendered in typing the manuscript.

ABSTRACT

This thesis carries a title that might appear to be too extensive as a topic. However, those familiar with the literature on biased estimators may agree that there is a well defined class of estimation procedures of interest to both mathematical statisticians and econometricians.

Efforts to introduce ideas which deviate from the traditional classical notion of unbiasedness have encountered enormous resistance. Admittedly, results relating to biased estimators are not as well-established as those relating to unbiased estimators, but unbiasedness is an arbitrary and unnecessarily stringent criterion. One should not therefore neglect the usefulness of biased estimators. With this background in mind, the thesis was written to synthesize the many differently motivated contributions which aim at improved estimation of unknown economic linear relationships. Apart from highlighting the author's own contributions in the area, the author has also attempted to make the thesis a self-contained one.

Chapter 1 motivates the study and defines the framework in which new estimators are developed. The fundamentals of Bayesian inference are discussed and the relation between formal and empirical Bayes procedures is examined. Chapter 2 provides a synthesis of different attempts to improve upon the traditional unbiased estimator. This chapter is necessary because it is not generally acknowledged that the differently motivated efforts can lead to the same result - namely, some sort of shrinkage must be introduced to improve estimation and that all the improved estimators are basically generalised Bayes rules. Chapter 3 introduces the controversial ridge estimator and provides a comprehensive survey. A new contribution made in this chapter is the introduction of a recursive algorithm for generating the ridge trace.

Chapter 4, 5 and 6 form the core of the thesis where new ideas are developed. Specifically, Chapter 4 attempts theoretical and Monte Carlo studies of the potential and realised reduction in risk of the biased estimators. A number of good adaptive ridge estimators are identified. As an illustration these are applied to re-estimating an investment function. Significantly more accurate predictions are achieved by the biased estimators than by conventional ordinary least squares estimator and the preliminary test estimators. Two new contributions are made in Chapter 5. Firstly, an analysis of seasonal variability in the distributed lag model sets the stage for the introduction of various estimators which can incorporate bi-dimensional prior information in the form of exchangeability and smoothness. Secondly, estimation of distributed lag model in the frequency domain is justified and the Spectral Ridge Estimator is introduced as an extension of Hannan's Efficient Estimator. The estimator's performance is compared to other well-known estimators using Almon's data. Chapter 6 works out the small sample bias and mean square error of a Generalised Ridge Instrumental Variable estimator for a structural equation in the context of a simultaneous equation system. The problem of undersized sample is tackled and the traditional optimism about 2SPC questioned. A new estimator which involves the application of ridge regression instead of the traditional least square regression at both stages of a 2SLS procedure is proposed and its statistical properties analysed (both asymptotically and in finite sample). Some further results concerning ridge regression are presented in the last chapter, i.e. 7. The robustness of ridge regression under misspecification is analysed. Problems of testing stochastic hypotheses and the construction of confidence sets are also discussed. Some of the criticisms of the technique are reviewed and a personal view is expressed.

TABLE OF CONTENTS

CHAPTER 1

AN OVERVIEW AND PRELIMINARIES

§1.1 Introduction

Few econometricians would take a particular set of coefficient estimates in a regression analysis too seriously. If the numbers are "good" they would have investigated further and tried to reconcile them with results obtained independently by other researchers. On the other hand, if the results appear to be unreasonable, usually they would suspect the results and follow up with a respecification of the model or selection of new proxies to replace the offending variables. This informal utilization of a prior information has jeopardized the reputation of econometrics as a 'science'. The value laden terms such as 'data mining', 'fishing' or 'data massaging' are often used to degrade each other's empirical work. Some econometricians of the Neyman-Pearson school have developed numerous statistical tests just to help carry out such activities. Two most commonly used strategies, which are dual to one another, have appeared in the literature in various guises. Simply, one extreme is to start searching from a very general model which presumably includes the true model. A series of Wald-type hypothesis tests are carried out to simplify the large model until further restrictions are rejected. This approach was advocated by Anderson (1971) and Mizon (1977) because the tests are uniformly most powerful and are independent of one another if the maintained hypotheses can be properly nested. The other extreme is to start from a simple but computationally convenient model and then apply the so-called Lagrange-multiplier type tests to see if the model need be generalised. This technique lacks the properties of being uniformly most powerful and of independence of the

Wald-type test but is a convenient diagnostic device for checking model misspecification. It was advocated by Aitchison and Silvey (1958). Both of the techniques are designed to avoid the heavy computational requirement of the standard likelihood ratio tests. An interesting study of these testing procedures is in Breusch (1978).

If our objective were solely to arrive at an adequate model, the above diagnostic tests are indispensable. This is perhaps in line with the Popperian principle that any hypothesis that is not consistent with facts must not be accepted. However, if inference is to be of any value, then somehow it must be leading to some courses of action. One important task of the econometrician is to provide the best possible estimates of some unknown quantities of the model. Social phenomena are too intricate to model correctly. The observation taken may not reflect the behaviour of the actual variable we believe to be of interest. Most of the economic variables cannot be directly observed and we must unfortunately be content with proxy variables. Thus, we are inevitably working with a false model most of the time. This fact has led Theil (1971) to stress in his book that it requires maturity for the student to realize that models are to be used but not to be believed.

Unlike physicists and other exact scientists whose main aim is to discover universal laws, econometricians could only hope to help economists to make a 'rational' decision in the face of uncertainty. All rational decisions are by definition those maximising a utility function or minimizing a cost function. This thesis is a study of estimators that would hopefully help to make better decisions than conventional ones. The heavy dose of Bayesian argument may upset many a classicist. Nevertheless, the estimators proposed herein do not rely on any Bayesian principle of inference for their validity. Only that we have utilized

Bayes theorem to guide us in searching for rules that may work better on the average than those arising without their guidance. We take special note of the data problem in economics - namely, multicollinearity - which prevents us from estimating economic relations accurately.

§1.2 Purposes of the Research

The main purposes of this study are:-

(a) to review critically the conventional estimators in the light of new statistical discoveries;

(b) to critically examine the advantages and problems associated with the use of the improved estimators in econometric data analysis;

(c) to adapt and propose new estimators to suit econometric investigation;

(d) to develop operational procedures which require non-trivial modification of existing ones;

(e) to carefully work out the statistical implications of new estimators so as to compare analytically with the more familiar estimators;

(f) If analytical technique fails, Monte Carlo studies are carried out to provide some qualified evidences of the performance of the new estimators.

It is difficult to overemphasize the importance of biased estimators in econometric analysis and there remains plenty of unsolved problems for further research. The notable problems as yet to be solved are the determination of admissible confidence regions and hypothesis tests. Nevertheless, a better knowledge of biased point estimation is a

prerequisite to these problems and more importantly, that would help improve the quality of the estimates of economic quantities now commonly used in the profession.

§1.3 Data Analysis in Economics

All econometric research is based on a set of numerical data relating to certain economic variables, and makes inferences from the data about the ways in which these variables are related. An economic model is simply a set of assumptions about the nature of the relationships. In general, we summarize our perceptions about the phenomenon in the form of a model and our observations in the form of estimates. Mathematical statistics can only help to check if the model is consistent with the data. It is up to the economist to add "meanings" to the empirical evidences.

There have been different approaches to data analysis. The Box and Jenkins approach is possibly the best known. They do not attach much weight to discovering structures of particular 'black boxes'. Their approach to the problem of forecasting and control can be conveniently summarized as follows:-

1. From the interaction of theory and practice, a useful class of models for the purpose at hand is considered.
2. Because this class is too extensive to be conveniently fitted directly to data, rough methods for identifying subclasses of these models are developed, which agrees with the principle of parsimony.
3. The tentatively entertained model is fitted to data and its parameters estimated.

4. Diagnostic checks are applied with the object of uncovering possible lack of fit and diagnosing the cause.

5. Steps 2 to 4 are repeated until a satisfactory model emerges which is then used for forecasting and control.

What are the defects of this approach? It is not hard to see that there is an element of trial and error. Tests may be introduced in an ad hoc manner and the whole process is but a more efficient way to summarize a complex set of data.

Despite the rapid advances in economic and econometric theory, we are never really sure of what hypothesis to test. We are searching quantitative economic knowledge in the process of non-experimental model building and there are many admissible economic and statistical models which do not contradict our perceived knowledge of economic relations. Inevitably, a priori information is always called on to discriminate among competing models. However, informal use of prior information may give rise to inadmissible estimators. A convenient way to utilize a priori information which result in admissible estimators is through Bayes method.

The essence of Bayes method of data analysis is the Bayes rule of conditional probability

$$f(\theta|y) = \frac{f(y|\theta)f(\theta)}{f(y)} \tag{1.1}$$

where $f(y|\theta)$ is the likelihood at the data point y

$f(\theta)$ is the prior probability function allocated to all possible subclasses θ of the general class of model to be considered

$f(\theta|y)$ is the posterior probability function describing the revision of the prior probability on receipt of the data y.

This approach makes use of the decision maker's own prior information about θ and uses the argument that even in the absence of data, decision makers do make reasonable decisions. Since uncertain a priori information is already available before the data is observed, the decision maker's uncertainty can be summarized in a probability function on θ. Based on the observed data, the likelihood function can then be combined with the a priori density to yield the posterior density of θ, which is supposed to represent the decision maker's total knowledge of the parametric distribution. All subsequent estimation and inference is then based on this posterior density of θ subject to a set of basic principles of consistent behaviour.

Some objections to the Bayesian theory are directed at the notion of randomness of the parameters which are supposed to be fixed. However, the Bayesian analysis does not really regard θ as a random outcome of some actual experiment but only argues that people who wish to decide consistently in uncertain situations should act as though θ were a random variable with a certain distribution function.

The problem of specifying an appropriate prior density is a real and difficult one. Distinction should be made between informative and non-informative prior densities. The fact that it is difficult to find a family of prior densities that is rich enough to incorporate the decision maker's belief but simple enough to be algebraically workable had led Jeffreys (1961) to consider the non-informative diffuse prior density. It is noted by Hill (1974) that indiscriminate use of such a diffuse prior density would violate the principle of coherency of de Finetti (1974). As considerable prior information about parameters is usually available, informative priors such as the natural conjugate family of Raiffa and Schlaifer (1961) are intuitively appealing and mathematically convenient. Unfortunately, the natural conjugate theory

does not always give a convenient family of prior densities. For some econometric problems, the family is very restrictive and often will not fit the prior beliefs of the decision maker. Example like systems of simultaneous equations model, the conjugate family is not rich enough to incorporate the prior beliefs that economists typically possess [see Rothenberg (1973)]. If other priors are used, the marginal posterior density is generally not easy to handle analytically and the moments of the distribution must be determined by numerical methods which would be extremely complex for problems with a large number of parameters involved.

§1.4 Decision Theory and Estimation

In this section, we shall sketch the basic concepts of decision theory which are used to justify new estimators proposed in later chapters. We regard estimation as the "terminal decision" after a series of subsidiary decisions, such as data selection and choice of model, have been made. The general problem in statistical decision theory consists of a space Θ of possible states of nature and a space A of possible actions and a loss function $L(\theta,d)$ defined in $\{\Theta \times A\}$. The loss depends on the outcome y through the function d which is used to assign an action for a given y. The problem is to select an "appropriate" decision rule d. Suppose an estimator for the $p\times 1$ vector θ, call it $d(y)$, is selected the loss function $L(\theta,d)$ depends on the true parameter $\theta \in \Theta$ and the estimate $d \in \mathcal{D}$. Assume the expectation of $L(\theta,d)$ given θ exists the risk function is thus defined.

$$R(\theta,d) = \mathop{\mathcal{E}}_{\mathcal{D}}[L(\theta,d)] = \int_{\mathcal{D}} L(\theta,d)f(y|\theta)\,\partial y \qquad (1.2)$$

As the true value of θ is uncertain, the estimator of the parameter having a smaller risk for all values of θ is always preferable

to one that has a large risk. Unfortunately, usually the risk function depends on the true value of θ and there is no one single estimator d with minimum risk for all values of θ with respect to all loss functions. It would be impossible to find an optimal estimator without further restriction on the class of estimators to be considered. The classical approach imposes restrictions such as unbiasedness, sufficiency, invariance, minimum variance unbiasedness. Perhaps a more desirable property of an estimator is admissibility.

Definition 1.4.1 (Domination)

An estimator d_1 is said to dominate an estimator d_2 if for all $\theta \in \Theta$, $R(\theta,d_1) \leq R(\theta,d_2)$. If, in addition, strict inequality holds for some $\theta \in \Theta$, then d_1 is said to strictly dominate d_2.

Definition 1.4.2 (Admissibility)

An estimator d_o is said to be admissible if it is *not* strictly dominated by any other estimator. Otherwise it is said to be inadmissible.

Remarks: The idea of admissibility is important in a negative sense in that it enables certain rules to be excluded.

The usual restrictions on estimators are *linear* and *unbiased* which sometimes yield inadmissible estimators. Another popular principle is the conservative one of minimising the maxima of the risk function over the parameter space.

Definition 1.4.3 (Minimaxity)

An estimator d_o is said to be minimax within the class of estimator $\mathcal{D}$ if $d_o \in \mathcal{D}$ and

$$\sup_{\theta \in \Theta} R(\theta,d_o) \leq \sup_{\theta \in \Theta} R(\theta,d) \qquad \forall\ d \in \mathcal{D}$$

Remarks: While this is mathematically interesting as a way to choose a particular rule among the set of admissible rules, it unfortunately evades all reference to prior information, quantitative or qualitative. It is possible that d_o has worse property than d for all values of θ but has a slightly lower maximum. Also, there may be numerous estimators which yield the same $\sup_{\theta \in \Theta} R(\theta,d)$, hence minimax property is not unique. However, if we have some prior information about θ, representable in the form of a probability function, say $f(\theta)$, then an estimator defined directly in terms of the risk function is obtained.

Definition 1.4.4 (Bayes Rule)

An estimator d_B is said to be Bayes with respect to the distribution f on Θ if the expected value of the risk function with respect to the distribution f on Θ is minimum, i.e. for all other d in the class of randomised estimators that belong to $\mathcal{D}$.

$$R^*(d_B) = \mathcal{E}[R(\theta,d_B)] = \int_{\Theta}\int_{\mathcal{D}} L(\theta,d_B)f(y/\theta)f(\theta)\partial y\partial\theta \leq R^*(d) \ .$$

Since different estimators have optimal properties under different loss functions, there is no consensus on which particular loss function to be chosen. For many practical purposes in economics, we naturally require our loss function to be zero if the estimator error is zero and to be an increasing function of the magnitude of the error. Suppose we approximate the loss function by a second order Taylor series,

$$L(\theta,d) = f(d-\theta) \doteq f(0) + (\nabla f)'(d-\theta) + (d-\theta)'W(d-\theta)$$

where $f(\cdot)$ is assumed to be twice differentiable, and $\nabla f(\cdot)$ is the gradient of the function. W is of course the Hessian matrix.

The fact that $f(0) = 0$ and that $f(\cdot)$ is non-negative implies

that $(\nabla f) = 0$ and that W is non-negative definite. Consequently, we have the general quadratic loss function

$$L_W(\theta,d) = (d-\theta)'W(d-\theta) \tag{1.3}$$

and its expected value, the risk

$$R_W(\theta,d) = \underset{\mathcal{D}}{\mathcal{E}} L_W(\theta,d) = \underset{\mathcal{D}}{\mathcal{E}}(d-\theta)'W(d-\theta) \tag{1.4}$$

Theorem 1.4.1 (Minimum Average Risk Estimator)

Under the quadratic loss (1.3), the Bayes estimator with respect to a certain distribution for θ is the conditional expectation of θ, $d_B = \underset{\Theta}{\mathcal{E}}(\theta/y)$ i.e. the posterior mean.

Proof.

$$\underset{\Theta}{\mathcal{E}} R_W(\theta,d) = \underset{\Theta}{\mathcal{E}} \underset{\mathcal{D}}{\mathcal{E}}(d-d_B)'W(d-d_B) + \underset{\Theta}{\mathcal{E}} \underset{\mathcal{D}}{\mathcal{E}}(d_B-\theta)'W(d_B-\theta) \tag{1.5}$$

$$\geq \underset{\Theta}{\mathcal{E}} R_W(\theta,d_B)$$

equality holds if and only if $d = d_B$. □

Note that we have assumed all the expectations exist. If θ is unbounded, then the posterior distribution of θ has no mean so that there is no estimate for which the expected loss is finite. Even if the mean of the posterior distribution exists, the variance of that distribution may not nor will the Bayes risk.

Theorem 1.4.2 (Admissibility of Bayes Estimator under L_W)

Any Bayes estimator obtained by taking a *proper* prior distribution over the whole parameter space must be admissible.

Proof. Suppose the risk of the Bayes estimator d_B exceeds some other estimator d such that

$$R_W(\theta,d) \leq R_W(\theta,d_B) \tag{1.6}$$

with strict inequality on a parameter set of positive probability under the prior density. Then clearly,

$$R^*(d) = \underset{\Theta}{\mathcal{E}} R_W(\theta,d) < \underset{\Theta}{\mathcal{E}} R_W(\theta,d_B) = R^*(d_B) \tag{1.7}$$

which contradicts the fact that $R^*(d_B)$ is a minimum. □

Remark: If the prior distribution is improper and if (1.6) is satisfied with strict inequality on some bounded set, then (1.7) does not necessarily follow.

Two special quadratic loss functions call for special attention. If covariance matrix of an estimator d is Σ, then $W = \Sigma^{-1}$ is the well-known case of invariant loss function. If $W = I$, the loss function represents the Euclidean distance between the estimator vector and the parameter vector.

Another often used loss function is the error matrix being the squared difference between d and the true parameter vector θ. Let us denote the *error matrix* by

$$L_2 = (d-\theta)(d-\theta)' \tag{1.8}$$

and the *risk matrix measure*

$$MSE(\theta,d) = \underset{\mathcal{D}}{\mathcal{E}} L_2 = \text{var}(d) + (\text{bias}(d))(\text{bias}(d))' \,. \tag{1.9}$$

The diagonal elements are the individual mean square errors $\underset{\mathcal{D}}{\mathcal{E}}(d_i-\theta_i)^2$ $i = 1 \ldots p$ whose sum is equal to the risk under quadratic loss with

$W = I$, i.e., the total mean square error of d is

$$TMSE(d) = \text{tr}\, MSE(\theta,d) .$$

Definition 1.4.5 (Admissibility under L_2).

An estimator d_2 is said to be admissible for θ under the loss (1.8) if there exists no other estimator d such that

$$MSE(\theta,d_2) - MSE(\theta,d) \qquad (1.10)$$

is non-negative definite and is positive definite for at least one value of θ.

Remarks: If d is admissible for θ under L_2, then $p'd$ may not be admissible for $p'\theta$ under mean square error for every p i.e. d may not be admissible for θ under quadratic loss with $W = pp'$. However, if d is admissible for θ under quadratic loss with $W = pp'$ for every p, then d is admissible for θ under L_2.

Theorem 1.4.3 (Shinozaki)

If d is admissible for θ under L_W for some positive definite matrix $W = W_o$, then d is admissible under L_W for any non-negative definite matrix W.

[A proof can be found in Rao (1976)].

Remarks: This theorem is useful in the sense that when comparing different estimators, condition for d to be admissible for θ under the quadratic loss with $W = I_p$ (i.e. L_I) is sufficient to guarantee admissibility of d for θ under the quadratic loss L_W for any non-negative definite matrix W.

Some econometricians may object to the use of simple quadratic loss and may also argue that the properties of decision rules are irrelevant

as far as inference is concerned. It should be pointed out that conventional econometric analysis rarely justifies the methods used on the basis of their value for economic decision making. The losses involved in using alternative statistical techniques are never considered. Although it is difficult to have a loss function to suit all problems, the quadratic loss is a reasonable approximation for symmetric problems. Sargan (1973) gave an account of the relative merits of the Cohen type and Bayesian type loss functions. The choice of the quadratic loss is not as arbitrary as the restriction of unbiasedness as the latter could be totally irrelevant for decision making. Although it often happens that the classical point estimates are similar to the optimal decision rules, the Bayesian decision-theoretic approach has the advantage that it can formally incorporate imprecise knowledge about the unknown parameters and at the same time places due attention on the loss functions being considered. We shall now turn to a brief survey of some standard results in Bayesian inference.

§1.5 Some Results in Bayesian Inference.

We have seen in the last section that given a parameter vector θ and a loss function $L(\theta,d)$, the best decision function d is the one with smallest expected loss (or risk). The statistical decision problem is to find the best decision when θ is unknown and there is, in general, no decision rule which has uniformly minimum risk. The traditional solution of restricting the class of decision rules to that of unbiased or that of minimax has been discussed. Here we shall be concerned with the so-called Bayesian solution which is based on the assumption that the decision maker's uncertainty can be summarized by a probability function on θ. The Bayesian solution to the statistical decision problem is to select the act that minimises expected loss where the expectation runs

over the joint distribution of y and θ. In practice, the Bayes risk at Definition 1.4.4 is not the most convenient one for analyzing the data. A simpler and equivalent approach is by invoking the Bayes theorem which relates the posterior density to the likelihood and the prior density as in (1.1). Thus, the Bayes risk can be rewritten as

$$R^*(d) = \int_{\mathcal{D}} \left[\int_{\Theta} L(\theta,d) f(\theta/y) \partial\theta \right] f(y) \partial y \ .$$

Since $f(y)$ is non-negative, $R^*(d)$ will be minimised if for every y,

$$\int_{\Theta} L(\theta,d) f(\theta/y) \partial\theta$$

is minimised. Hence the optimal Bayes decision is that which minimises expected posterior loss.

Restricting our attention to quadratic loss, we have the result from Theorem 1.4.1 that the posterior mean is the best Bayes decision rule. In what follows, the posterior means corresponding to various prior densities of the parameters for the standard linear regression model will be exhibited. Other results of Bayesian inference which are beyond the scope of this thesis could be found in Zellner (1971).

Consider now the normal multiple regression model

$$y = X\beta + u \qquad (1.11)$$

where y is a column of T observations on the dependent variables. X is a $T\times p$ matrix of exogenous variables which are independent of the $(T\times 1)$ random vector u. β is $p\times 1$ vector of parameters we wish to estimate. Without loss of generality, we assume that the vector u is distributed as T-dimensional spherical normal with mean zero and

variance σ^2. Hence, the likelihood function of the unknowns β and σ^2 is

$$f(y|X,\beta,\sigma^2) \propto \frac{1}{\sigma^T} \exp - \frac{1}{2\sigma^2} [(\beta-\hat{\beta})'X'X(\beta-\hat{\beta}) + s^2] \qquad (1.12)$$

where

$$\hat{\beta} = (X'X)^{-1}X'y$$

$$s^2 = (y-X\hat{\beta})'(y-X\hat{\beta})$$

The likelihood is to be combined with whatever prior densities on (β,σ^2) to give the corresponding posterior densities. In general, the operation is not a simple one except for some of the special cases considered below:-

(a) Diffuse (non-informative) Prior Density on (β,σ^2)

If the decision maker has absolutely no information about the elements of β and σ, Jeffreys (1961) suggested that one can assume the elements of β and $(\log \sigma)$ as being independent and uniformly distributed.

That is

$$f(\beta,\sigma) \propto \frac{1}{\sigma} \qquad -\infty < \beta_i < \infty \quad i = 1 \ldots p \qquad (1.13)$$
$$0 < \sigma < \infty$$

Invoking Bayes theorem, we obtain the posterior density of (β,σ) given y as proportional to the product of (1.12) and (1.13)

$$f(\beta,\sigma|y) \propto \frac{1}{\sigma^{T+1}} \exp - \frac{1}{2\sigma^2} [s^2 + (\beta-\hat{\beta})'X'X(\beta-\hat{\beta})] \qquad (1.14)$$

which is easily recognised to be a p-dimentional normal distribution for β conditional on a value of σ, with mean $\hat{\beta}$ and covariance matrix $\sigma^2(X'X)^{-1}$. As σ^2 is rarely known, the nuisance parameter can be

eliminated by integrating (1.14) with respect to σ to obtain the marginal posterior density of β as

$$f(\beta|y) \propto \{s^2 + (\beta-\hat{\beta})'X'X(\beta-\hat{\beta})\}^{-\frac{T}{2}} \tag{1.15}$$

which is then the familiar multivariate Student-t distribution. The posterior density of σ can be obtained similarly, but we shall only consider the estimator of β.

(b) Natural Conjugate (informative) Prior Density on (β,σ^2)

Another type of prior that appears to be more reasonable than the diffuse and yet yields tractable posterior density was proposed by Raiffa and Schlaifer (1961). The "natural conjugate" family takes advantage of the idea that the product of two independent normal densities is also a normal density. So that with respect to the likelihood function at (1.12), the natural conjugate prior for (β,σ^{-2}) is the normal-gamma-2 density. This idea arises naturally from formulating the prior density based on a previous regression analysis. That is, the prior

$$f(\beta|\sigma^2) \propto \frac{1}{\sigma^p} \exp -\frac{1}{2\sigma^2}(\beta-\bar{\beta})'M(\beta-\bar{\beta}) \qquad -\infty < \beta < \infty \tag{1.16}$$

and

$$f\left(\frac{1}{\sigma^2}\right) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{\nu}{2}-1} e^{-\frac{\bar{s}^2}{2\sigma^2}} \qquad \bar{s}^2 > 0 \quad \nu > 0 \quad \sigma^2 > 0 \tag{1.17}$$

$$\left[\text{i.e.} \quad \beta \sim N_p(\bar{\beta},\sigma^2M^{-1}) \qquad \frac{\bar{s}^2}{\sigma^2} \sim \chi^2_\nu \right].$$

Combining, we have the joint normal-gamma-2 prior

$$f\left(\beta,\frac{1}{\sigma^2}\right) \propto \frac{1}{\sigma^{p+\nu-2}} \exp -\frac{1}{2\sigma^2}[(\beta-\bar{\beta})'M(\beta-\bar{\beta}) + \bar{s}^2] \qquad p+\nu \geq 0 . \tag{1.18}$$

Thus, the posterior density of (β,σ^2) is

$$f\left(\beta, \frac{1}{\sigma^2} \mid y\right) \propto \frac{1}{\sigma^{T+p+\nu-2}} \exp - \frac{1}{2\sigma^2} [(\beta-\beta^*)'M^*(\beta-\beta^*) + S^{*2}] \qquad (1.19)$$

where

$$M^* = X'X + M$$

$$\beta^* = M^{*-1}[X'X\hat{\beta}+M\bar{\beta}] \quad \text{and} \quad S^{*2} = y'y + \bar{s}^{-2} + \bar{\beta}'M\bar{\beta} - \beta^{*\prime}M^*\beta^* \ .$$

This is again a p-dimensional normal distribution for β with mean β^* and variance $\sigma^2 M^{*-1}$. Integrating out $\frac{1}{\sigma^2}$, we have the marginal posterior density of β in multivariate Student-t form

$$f(\beta|y) \propto [S^{*2} + (\beta-\beta^*)'M^*(\beta-\beta^*)]^{-\frac{T+p+\nu}{2}} \ . \qquad (1.20)$$

In the limiting case, where $p+\nu = 0$, $M = 0$, $\bar{s}^{-2} = 0$, and we have the (degenerate) diffuse prior density, then

$$f(\beta|y) \propto [s^2 + (\beta-\hat{\beta})'X'X(\beta-\hat{\beta})]^{-\frac{T}{2}} \ . \qquad (1.21)$$

This justifies the interpretation of the natural conjugate prior (1.18) as a posterior density of a previous regression based on a sample of size $p+\nu$ under a diffuse original prior density.

(c) Hyper-Parameter Models

As a further justification for the natural-conjugate family, Lindley and Smith (1972) considered the possibility of a probability distribution for the parameters of the prior distribution. Within the confines of the normal distribution, such hyper-parameter models are particularly simple as well as sensible. Suppose, given θ_1 and C_1

$$y \sim N_T(A_1\theta_1, C_1) \qquad (1.22)$$

and the prior distribution of θ_1 is, given θ_2, C_2,

$$\theta_1 \sim N_p(A_2\theta_2, C_2)$$

so that the posterior distribution of θ_1 is

$$\theta_1|y \sim N_p(Bb, B) \tag{1.23}$$

where

$$B^{-1} = A_1'C_1^{-1}A_1 + C_2^{-1}$$

$$b = A_1'C_1^{-1}y + C_2^{-1}A_2\theta_2 \ .$$

By considering θ_2 as random, normally distributed with mean $A_3\theta_3$ and covariance C_3, then the marginal prior distribution of θ_1, given θ_3 and C_3, is

$$\theta_1 \sim N(A_2A_3\theta_3, C_2 + A_2C_3A_2')$$

and using this as prior, the posterior density of θ_1 becomes

$$\theta_1|y \sim N(Dd, D) \tag{1.24}$$

where

$$D^{-1} = A_1'C_1^{-1}A_1 + (C_2 + A_2C_3A_2')^{-1}$$

$$d = A_1'C_1^{-1}y + (C_2 + A_2C_3A_2')^{-1}A_2A_3\theta_3 \ .$$

Lindley and Smith indicated that one can again impose a distribution on θ_3 and continue up to as many stages as one wishes but suggested that the process could be stopped at the second stage for many practical purposes.

In the limiting case, when the second stage prior is vague, i.e. $C_3 \to \infty$, then using the identity (see Rao (1973) p.33)

$$(C_2 + A_2C_3A_2')^{-1} = C_2^{-1} - C_2^{-1}A_2(A_2'C_2^{-1}A_2 + C_3^{-1})^{-1}A_2'C_2^{-1} \ .$$

and setting $C_3^{-1} = 0$ so that $(C_2 + A_2 C_3 A_2')^{-1} A_2 = 0$, we obtain

$$\theta_1 | y \sim N(D_o d_o, D_o) \tag{1.25}$$

where

$$D_o^{-1} = A_1' C_1^{-1} A_1 + C_2^{-1} - C_2^{-1} A_2 (A_2' C_2^{-1} A_2)^{-1} A_2' C_2^{-1}$$

$$d_o = A_1' C_1^{-1} y \ .$$

This result can be applied to the linear regression model. In particular, if the assumption of exchangeability is plausible such that the elements of β are independent and as important as each other, then the prior for β is $N(\mu \underset{\sim}{\ell}_p, \sigma_\beta^2 I_p)$ *where* $\underset{\sim}{\ell}_p$ is a column of p units and μ is a scalar. If we further assume that the information about μ is vague, then the posterior distribution of θ_1 is

$$\theta_1 | y \sim N(D_o d_o, D_o)$$

where

$$D_o^{-1} = \frac{1}{\sigma^2} X'X + \frac{1}{\sigma_\beta^2} [I_p - \frac{1}{p} J_p]$$

$$d_o = \frac{1}{\sigma^2} X'y$$

$$J_p = \ell_p \ell_p' = \text{a } p \times p \text{ matrix of units.}$$

It is easily seen that the posterior mean of β

$$\beta^* = D_o d_o = [X'X + \frac{\sigma^2}{\sigma_\beta^2} (I_p - \frac{1}{p} J_p)]^{-1} X'y \tag{1.26}$$

is a form of ridge-type regression which we shall discuss in Chapter 3.

The same idea can be trivially extended to a system of m seemingly unrelated regressions *where*

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}$$ is a stacked vector of the m (T×1) vectors.

$$A_1 = \begin{pmatrix} X_1 & & 0 \\ & \ddots & \\ 0 & & X_m \end{pmatrix}$$

$$\theta_1 = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}$$

$$C_1 = \begin{pmatrix} \sigma_1^2 I_T & & \bigcirc \\ & \ddots & \\ \bigcirc & & \sigma_m^2 I_T \end{pmatrix} = \Omega \otimes I_T$$

$$\Omega = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_m^2 \end{pmatrix} .$$

If we impose an exchangeable prior for β_j, say, $N_p(\mu,\Sigma)$ $j = 1 \ldots m$, i.e. $\theta_1 \sim N(A_2\theta_2, C_2)$ with

$$A_2 = \ell_m \otimes I_p$$

$$\theta_2 = \mu \text{ , a } p\times 1 \text{ vector}$$

$$C_2 = I_m \otimes \Sigma \ .$$

Introducing a diffuse prior on θ_2, we can use the result at (1.24) to arrive at a posterior mean for β_j as

$$\beta_j^* = \left(\frac{1}{\sigma_j^2} X_j'X_j + \Sigma^{-1}\right)^{-1}\left(\frac{1}{\sigma_j^2} X_j'y_j + \Sigma^{-1}\ \bar{\beta}^*\right) \qquad j = 1 \ldots m \qquad (1.27)$$

where

$$\bar{\beta}^* = \sum_{j=1}^{m} w_j \hat{\beta}_j$$

$$\hat{\beta}_j = (X_j'X_j)^{-1}X_j'y_j$$

$$w_j = \left[\sum_{i=1}^{m} \left(\frac{X_i'X_i}{\sigma_i^2} + \Sigma^{-1} \right)^{-1} \frac{X_i'X_i}{\sigma_i^2} \right]^{-1} \left(\frac{X_j'X_j}{\sigma_j^2} + \Sigma^{-1} \right)^{-1} \frac{X_j'X_j}{\sigma_j^2} \quad .$$

(d) Poly-t Densities

Poly-t densities defined in Drèze (1977) are those having their kernel as a product, or ratio of products, of Student-t kernels. They are obtained as posterior marginal densities for regression coefficients, under a variety of specifications for the prior density and the data generating process. In this thesis, we shall have occasion to use only the product form of poly-t density and hence only this form will be discussed.

Suppose that two independent normal samples, labelled 1 and 2, yield respectively T_1 and T_2 observations on the linear model (1.11). $\mathcal{E}u_1 = \mathcal{E}u_2 = 0$, $\mathcal{E}u_1u_1' = \sigma_1^2 I_{T_1}$, and $\mathcal{E}u_2u_2' = \sigma_2^2 I_{T_2}$, $\sigma_1^2 \neq \sigma_2^2$. The joint density of the combined sample is

$$f(y_1,y_2|\beta,\sigma_1^2,\sigma_2^2) \propto \frac{1}{\sigma_1^{T_1}} \cdot \frac{1}{\sigma_2^{T_2}}$$

$$\exp -\frac{1}{2} \left[\frac{(y_1-X_1\beta)'(y_1-X_1\beta)}{\sigma_1^2} + \frac{(y_2-X_2\beta)'(y_2-X_2\beta)}{\sigma_2^2} \right] . \qquad (1.28)$$

Using the diffuse prior

$$f(\beta,\sigma_1^{-2},\sigma_2^{-2}) \propto \frac{1}{\sigma_1^2} \frac{1}{\sigma_2^2} \qquad (1.29)$$

we obtain the posterior densities

$$f(\beta,\sigma_1^2,\sigma_2^2|y_1,y_2) \propto \frac{1}{\sigma_1^{T_1-2}} \cdot \frac{1}{\sigma_2^{T_2-2}}$$

$$\exp -\frac{1}{2}\left[\frac{(y_1-X_1\beta)'(y_1-X_1\beta)}{\sigma_1^2} + \frac{(y_2-X_2\beta)'(y_2-X_2\beta)}{\sigma_2^2}\right] \tag{1.30}$$

Now, integrating with respect to σ_1^{-2} and σ_2^{-2}, we have the marginal posterior density of β as

$$f(\beta|y_1,y_2) \propto$$

$$[s_1^2 + (\beta-\hat{\beta}_1)'X_1'X_1(\beta-\hat{\beta}_1)]^{-\frac{T_1}{2}} [s_2^2 + (\beta-\hat{\beta}_2)'X_2'X_2(\beta-\hat{\beta}_2)]^{-\frac{T_2}{2}} \tag{1.31}$$

where $n_i = T_i-p$ $(i = 1,2)$ and $s_i^2 = (y_i-X_i\hat{\beta}_i)'(y_i-X_i\hat{\beta}_i) \sim \sigma_i^2\chi_{n_i}$, which is seen to be a product of two multivariate Student-t kernels.

Another way for such a poly-t form to arise is when we are using a priori density which is Student type. This arises naturally from the Lindley-Smith hyper-parameter framework where the location parameter of the prior normal distribution for β has a diffuse second stage prior while the scale parameter is a gamma-2 type. Then the marginal prior distribution for β is a multivariate Student-t distribution being regarded as the marginal posterior distribution from an original diffuse prior. That is,

$$\hat{\beta} \sim N_T(\beta,\sigma^2(X'X)^{-1}) \ . \tag{1.32}$$

We already know that if $f(\beta,\sigma^2) \propto \frac{1}{\sigma^2}$, then the posterior density is

$$f(\beta,\sigma^2|\hat{\beta}) \propto \frac{1}{\sigma^{T+2}} \exp -\frac{1}{2\sigma^2}[s^2 + (\beta-\hat{\beta})'X'X(\beta-\hat{\beta})]$$

and

$$f(\beta|\hat{\beta}) \propto [s^2 + (\beta-\hat{\beta})'X'X(\beta-\hat{\beta})]^{-\frac{T}{2}} ,$$

where

$$\frac{s^2}{\sigma^2} \sim \chi^2_{T-p} .$$

So, by the same argument if $\beta \sim N_p(\theta,\tau^2M^{-1})$ and if $f(\theta,\tau^2) \propto \frac{1}{\tau^2}$, then the marginal posterior distribution of θ is

$$f(\theta|\beta) \propto [\bar{s}^2 + (\theta-\beta)'M(\theta-\beta)]^{-\frac{p}{2}} \tag{1.33}$$

where

$$\frac{\bar{s}^2}{\tau^2} \sim \chi^2_{\nu} .$$

Now if we take the prior for (β,σ^2) as the product of a Student density for β and a uniform distribution for σ^{-2}, then

$$f(\beta,\sigma^{-2}) \propto \frac{1}{\sigma^2} [\bar{s}^2 + (\theta-\beta)'M(\theta-\beta)]^{-\frac{p}{2}} . \tag{1.34}$$

Combining (1.34) with the likelihood function and integrating with respect to σ^{-2} we have

$$f(\beta|y) \propto [\bar{s}^2 + (\theta-\beta)'M(\theta-\beta)]^{-\frac{p}{2}} \left| [s^2 + (\beta-\hat{\beta})'X'X(\beta-\hat{\beta})]^{-\frac{T}{2}} \right. \tag{1.35}$$

which is a rather complicated function and numerical integration is generally required to obtain its moments. In view of this difficulty, Dickey (1975) and Leamer (1969) considered the computation of modal values instead of the posterior mean. An alternative approach adopted by Tiao and Zellner (1964) is to expand the poly-t density in a Taylor

series and compute approximate moments from the leading terms of the series. The first term, being the kernel of a normal density, is usually of interest. Drèze (1977) cautioned that poly-t densities are typically asymmetrical and multimodal, so that the aforementioned approximations are valid only in special cases.

§1.6 Empirical Bayes Procedures

The use of normal priors with normal likelihood function results in normal posterior distributions. They are mathematically convenient at the cost of some restrictive assumptions that the nuisance parameters are known. As we have seen, the removal of these nuisance parameters by integration may lead to an intractable marginal posterior distribution. A convenient solution is to replace these nuisance parameters by some consistent estimates as was done in Rao (1975) and Efron and Morris (1973). The posterior mean so obtained is generally known as the Empirical Bayes Estimator. Apart from the problem of unknown nuisance parameters, the functional form of the prior distributions adopted by the usual Bayes procedure need be justified. One way to avoid these criticisms is to follow the non-parametric empirical Bayes procedure of Robbins (1956). Appreciable work of this type relating to multiple regression has been done in the literature [see Martz and Krutchkoff (1969)].

Consider again the normal regression model. It is well known that $\hat{\beta}$ and s^2 are jointly sufficient such that

$$\hat{\beta} \sim N_p(\beta, \sigma^2(X'X)^{-1})$$

and independent of

$$s^2 \sim \sigma^2 \chi^2_{T-p} \,. \tag{1.36}$$

If we denote these distributions by $f(\hat{\beta}|\beta,\sigma^2)$ and $f_{T-p}(s^2|\sigma^2)$,

the joint density of $\hat{\beta}$ and s^2 can be written as

$$f_{T-p}(\hat{\beta},s^2|\beta,\sigma^2) = f(\hat{\beta}|\beta,\sigma^2)f_{T-p}(s^2|\sigma^2) \ . \tag{1.37}$$

The posterior mean of β based on some general prior distribution $G(\beta,\sigma^2)$ can be written as

$$\begin{aligned}\beta^* &= \&(\beta|\hat{\beta},s^2) \\ &= \int \beta f(\beta,\sigma^2|\hat{\beta},s^2)d\beta d\sigma^2 \\ &= \int \beta \frac{f_{T-p}(\hat{\beta},s^2|\beta,\sigma^2)dG(\beta,\sigma^2)}{f_{T-p}(\hat{\beta},s^2)}\end{aligned} \tag{1.38}$$

where $f(\beta,\sigma^2|\hat{\beta},s^2)$ is the posterior distribution of β while the third equality was obtained by invoking Bayes Theorem.

From the normality of $\hat{\beta}$, we see that

$$\ell nf(\hat{\beta}|\beta,\sigma^2) \propto -\frac{1}{2\sigma^2}(\hat{\beta}-\beta)'X'X(\hat{\beta}-\beta) \tag{1.39}$$

and it follows that

$$\beta = \hat{\beta} + \sigma^2(X'X)^{-1}\frac{\partial \ell nf(\hat{\beta}|\beta,\sigma^2)}{\partial\hat{\beta}} \ . \tag{1.40}$$

Substituting (1.40) into (1.38) gives

$$\beta^* = \hat{\beta} + \frac{(X'X)^{-1}}{f_{T-p}(\hat{\beta},s^2)}\int \sigma^2 \frac{\partial \ell nf(\hat{\beta}|\beta,\sigma^2)}{\partial\hat{\beta}} f_{T-p}(\hat{\beta},s^2|\beta,\sigma^2)dG(\beta,\sigma^2) \ . \tag{1.41}$$

Using the fact that

$$\sigma^2 f_n(s^2|\sigma^2) = \frac{s^2}{n-2} f_{n-2}(s^2|\sigma^2) \ , \tag{1.42}$$

the integrand in (1.41) can be written as

$$\sigma^2 \frac{\partial f(\hat{\beta}|\beta,\sigma^2)}{\partial\hat{\beta} f(\hat{\beta}|\beta,\sigma^2)} \cdot f_{T-p}(\hat{\beta},s^2|\beta,\sigma^2)$$

$$= \sigma^2 \frac{f_{T-p}(\hat{\beta},s^2|\beta,\sigma^2)}{f(\hat{\beta}|\beta,\sigma^2)} \cdot \frac{\partial f(\hat{\beta}|\beta,\sigma^2)}{\partial\hat{\beta}}$$

$$= \sigma^2 f_{T-p}(s^2|\beta,\sigma^2) \cdot \frac{\partial f(\hat{\beta}|\beta,\sigma^2)}{\partial\hat{\beta}}$$

$$= \frac{s^2}{T-p-2} f_{T-p-2}(s^2|\beta,\sigma^2) \frac{\partial f(\hat{\beta}|\beta,\sigma^2)}{\partial\hat{\beta}}$$

$$= \frac{s^2}{T-p-2} \frac{\partial f_{T-p-2}(\hat{\beta},s^2|\beta,\sigma^2)}{\partial\hat{\beta}} .$$

Plugging back into (1.41), we have

$$\beta^* = \hat{\beta} + (X'X)^{-1} \frac{s^2}{T-p-2} \cdot \frac{\partial f_{T-p-2}(\hat{\beta},s^2)}{f_{T-p}(\hat{\beta},s^2)\partial\hat{\beta}} \qquad (1.43)$$

From this expression, it is clear that if $\hat{\beta}$ is the local maximum of the marginal distribution $f_{T-p-2}(\hat{\beta},s^2)$, then the derivative in (1.43) vanishes and the posterior mean $\beta^* \equiv \hat{\beta}$. If β_o ($\neq \hat{\beta}$) is the local maximum, then β^* will shrink $\hat{\beta}$ towards β_o [see Andrews et al (1972)]. The form in (1.43) is still not operational as the marginal distribution and its derivatives are usually unknown. However, consistent estimates for a density and its derivatives can be used [see Bhattacharya (1967) and Singh (1977)].

In the case of a known prior distribution, the posterior mean can be retrieved from (1.43). As the Jeffreys' prior specifies that $g(\beta,\sigma^2) \propto \frac{1}{\sigma^2}$, the marginal density of $\hat{\beta}$ becomes

$$f(\hat{\beta}) = \int f(\hat{\beta}|\beta,\sigma^2) g(\beta,\sigma^2) d\beta d\sigma^2$$

$$= \int \frac{1}{\sigma^2} d\sigma^2 = 1 \qquad (1.44)$$

Hence, $\frac{\partial f(\hat{\beta})}{\partial \hat{\beta}} = 0$ implying that $\beta^* = \hat{\beta}$.

On the other hand, if given σ^2, the prior distribution of β is $N_p(\theta, \tau^2 M^{-1})$, then the marginal distribution of $\hat{\beta}$ is $N_p(\theta, \sigma^2(X'X)^{-1} + \tau^2 M^{-1})$. Thus

$$\frac{\partial f(\hat{\beta})}{f(\hat{\beta})\partial \hat{\beta}} = \frac{\partial \ell n f(\hat{\beta})}{\partial \hat{\beta}} = -\frac{X'X}{\sigma^2}\left(\frac{X'X}{\sigma^2} + \frac{M}{\tau^2}\right)^{-1}\frac{M}{\tau^2}(\hat{\beta}-\theta) \tag{1.45}$$

substituting this into (1.40), we have

$$\begin{aligned}\beta^* &= \hat{\beta} + \sigma^2(X'X)^{-1}\frac{\partial f(\hat{\beta})}{f(\hat{\beta})\partial \hat{\beta}} \\ &= \hat{\beta} - \left(\frac{X'X}{\sigma^2} + \frac{M}{\tau^2}\right)^{-1}\frac{M}{\tau^2}(\hat{\beta}-\theta) \\ &= \left(\frac{X'X}{\sigma^2} + \frac{M}{\tau^2}\right)^{-1}\left(\frac{X'X}{\sigma^2}\hat{\beta} + \frac{M}{\tau^2}\theta\right) \qquad (1.46)\end{aligned}$$

which is in exactly the same form as if we had followed the conventional approach.

If the prior were a multivariate Student's t-distribution, the marginal distribution being a product of normal and Student kernels would result in a very complicated expression. The posterior mean could not be expressed in closed form and numerical method is usually required.

CHAPTER 2

CLASSICAL AND IMPROVED ESTIMATORS OF THE LINEAR MODEL

§2.1 The Classical Procedures

Econometric analysis begins with a linear model stating an approximate relationship of a dependent variable y_t with a weighted sum of p other explanatory variables $x_{1t} \ldots x_{pt}$ (one of which may be constant or dummy variable), the subscript t stands for the time point at which the variables are observed. We write a sample of T observations on the $p + 1$ variables compactly in matrix notation

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

where X is $T \times p$ nonstochastic and contemporarily uncorrelated with the random error u. β is $p \times 1$ vector of unknown constant weights which we want to estimate. Models of the form (2.1) are fundamental in econometrics: the single equation model with X as the exogenous variables in which the random errors may be serially correlated or heteroscedastic; the vector form of a system of seemingly unrelated regressions or the linearized version of a model which is intrinsically non-linear in parameters. Finally, if X is correlated with the random error as in the case of interdependent equations system or of the dynamic single equation model with both lagged dependent variables and serially correlated errors present, one can arrive at the form (2.1) through pre-multiplication by a set of instrumental variables which are known to be uncorrelated with the random errors but are good proxies for the X. With no information other than that the first moment of u is zero and the second moment is a positive definite symmetric matrix $\sigma^2 V$ (σ^2 is unknown), the classical method of estimating β is to minimise the following quadratic form with respect to β

$$(y - X\beta)'V^{-1}(y - X\beta) \tag{2.2}$$

leading to the so-called normal equations

$$X'V^{-1}X\hat{\beta} = X'V^{-1}y \ . \tag{2.3}$$

The solution of this normal equations, $\hat{\beta}$, is unique if the matrix $X'V^{-1}X$ is non-singular - this will be so if X is of full column rank p since V was assumed to be positive definite. Although the explanatory variables are, in general, highly correlated with one another in econometrics, the assumption that X is of rank p would normally be satisfied. The Generalised Least Square estimator (GLS) $\hat{\beta}$ is well known to be linear, unbiased and efficient such that any linear unbiased estimator has a variance matrix which exceeds that of $\hat{\beta}$ by a positive semidefinite matrix. In symbols, the GLS

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y \tag{2.4}$$

has mean

$$\mathrm{E}\hat{\beta} = \beta \tag{2.5}$$

and variance

$$\mathrm{var}\ \hat{\beta} = \sigma^2(X'V^{-1}X)^{-1} \ . \tag{2.6}$$

Furthermore, when the random error is normally distributed, the properties of GLS are overwhelming. It can be shown that the maximum likelihood estimator which attains the Cramer-Rao bound in this situation is identical to the GLS. This implies that GLS dominates the whole class of unbiased estimators (linear or not) under normality.

If the restriction of unbiasedness is relaxed, then small variance is not sufficient to ensure a good estimator since there exist estimators having zero variance but an infinitely large bias. The risk of the

resulting estimator would then depend on the unknown parameter. However, if a priori information are available about the possible location of the unknown parameters, then it is very likely that one is able to choose a linear but biased estimator that dominates GLS even under normality.

§2.2 Consequence of Misspecification

On many occasions, one would have applied the Least Squares method to a wrongly specified model since there exist many plausible or theoretically acceptable models that economic data are not sharp enough to discriminate among them. We shall primarily be concerned with the effect of misspecification on the risk of the GLS.

Suppose our true model is given in (2.1) and we think that more information will be available by including further variables (even if they are irrelevant). The model to be estimated is then

$$y = X\beta + Z\gamma + u \tag{2.7}$$

where Z is $T\times q$ extra variables. Denote $M_z = V^{-1} - V^{-1}Z(Z'V^{-1}Z)^{-1}Z'V^{-1}$. We can write the GLS of β on the model (2.7) as

$$\begin{aligned}\tilde{\beta} &= (X'M_zX)^{-1}X'M_zy \\ &= \beta + (X'M_zX)^{-1}X'M_zu \quad \because \quad M_zZ = 0 = Z'M_z \ .\end{aligned}$$

The mean and variance of $\tilde{\beta}$ are respectively

$$\mathcal{E}\tilde{\beta} = \beta$$

and

$$\text{var}\ \tilde{\beta} = \sigma^2(X'M_zX)^{-1} \quad \because \quad M_zVM_z = M_z \ .$$

Now, it is clear that $\text{var}(\tilde{\beta}) - \text{var}(\hat{\beta})$ is positive semidefinite because $X'V^{-1}X - X'M_zX = X'V^{-1}Z(Z'V^{-1}Z)^{-1}Z'V^{-1}X$ is positive semidefinite.

It follows that $\tilde{\beta}$ is unbiased but less efficient than $\hat{\beta}$. From a variance reduction point of view, it could only do worse by including irrelevant variables. Of course, if theory suggests that the weights of Z are of independent interest and should be estimated, then the true model is (2.7) rather than (2.1) and this problem belongs to the domain of model selection.

One common principle in econometric model building is that of *parsimony*. After all the possible explanatory variables are experimented with, the econometrician is tempted to drop variables that are not statistically significant or to impose some restrictions on the estimates in the hope that the resulting estimates may be "improved". This operation is what Leamer (1978) termed "Simplification Search". One usual motivation for dropping statistically insignificant variables is to correct the 'wrong' signs of the coefficient estimates of some 'important' variables. The futility of such effort is evident from the following results.

Theorem 2.2.1 (Visco (1978))

Let t_j be the t-ratio for the testing of the hypothesis that $\beta_j = 0$ based on the unrestricted estimate of β_j, $\hat{\beta}_j$ say. Let ρ_{ij} be the correlation coefficient between $\hat{\beta}_i$ and $\hat{\beta}_j$. Suppose that the hypothesis is accepted because $|t_i|$ is smaller than some prescribed critical value, then the necessary and sufficient condition that $\hat{\beta}_j$ and $\tilde{\beta}_j$ (the restricted estimate) will have opposite sign is

$$t_j^2 < \rho_{ij} t_i t_j \ . \tag{2.8}$$

Note that if $|t_j| > |t_i|$, then (2.8) can never be satisfied because $|\rho_{ij}| < 1$.

COROLLARY 2.2.1 (Leamer (1975))

There can be no change in sign of any coefficient that is more significant than the coefficient of the omitted variable. More generally, the restricted estimate of a parameter β_j must lie in the interval

$$[\hat{\beta}_j - \sqrt{var(\hat{\beta}_j)}|t_i| \ , \ \hat{\beta}_j + \sqrt{var(\hat{\beta}_j)}|t_i|]$$

(Proofs are in the referenced articles).

The result of Leamer can be generalised to the case when more complicated restrictions than the simple exclusion of variable are imposed. Suppose m linear independent restrictions on the parameters in the form of $H\beta = h$ are introduced. H is $m \times p$ of rank m and h is $m \times 1$, both are known nonstochastic quantities. The restricted estimate for a particular linear combination of the parameters $a'\beta$ will fall in the region (assuming without loss of generality that $V = I_T$, $\hat{\beta}$ is simply the Ordinary Least Square (OLS) estimator).

$$a'\hat{\beta} \pm \{a'(X'X)^{-1}a\hat{\sigma}^2 mF\}^{\frac{1}{2}}$$

where $F = \dfrac{(\tilde{\beta}-\hat{\beta})'X'X(\tilde{\beta}-\hat{\beta})}{m\hat{\sigma}^2}$ is the statistic for testing if $\hat{\beta}$ satisfies the restrictions, and $\hat{\sigma}^2 = \dfrac{(y-X\hat{\beta})'(y-X\hat{\beta})}{T-p}$. $\tilde{\beta}$ is the Restricted Least Square (RLS) estimator so that $H\tilde{\beta} = h$.

Proof. The restricted estimator $\tilde{\beta}$ of β subject to $H\beta = h$ is related to the unrestricted estimator $\hat{\beta}$ as

$$\tilde{\beta} = \hat{\beta} - (X'X)^{-1}H'[H(X'X)^{-1}H']^{-1}[H\hat{\beta}-h] \ .$$

That is $\tilde{\beta}$ lies on the ellipsoid

$$(\tilde{\beta}-\hat{\beta})'X'X(\tilde{\beta}-\hat{\beta}) = (H\hat{\beta}-h)'[H(X'X)^{-1}H']^{-1}(H\hat{\beta}-h)$$

$$= m\hat{\sigma}^2 F \ .$$

But Cauchy's inequality ensures that

$$|a'(\tilde{\beta}-\hat{\beta})|^2 \leq \{a'(X'X)^{-1}a\}\{(\tilde{\beta}-\hat{\beta})'X'X(\tilde{\beta}-\hat{\beta})\} .$$

Thus

$$-\{a'(X'X)^{-1}a\hat{\sigma}^2mF\}^{\frac{1}{2}} \leq a'(\tilde{\beta}-\hat{\beta}) \leq \{a'(X'X)^{-1}a\hat{\sigma}^2mF\}^{\frac{1}{2}}$$

Q.E.D. □

when $m = 1$ and $a = e_i$, the i'th column of I_p, Leamer's result becomes a special case.

Another motivation for imposing hypothetical restrictions is to reduce the variance of the coefficient estimates. It is well known that if the restrictions are correct, the restricted estimator is unbiased and has smaller variance than the unrestricted estimator. However, if the restrictions are incorrect, the restricted estimator is biased but a reduction in variance is still possible. If estimation is our goal, then consideration of risk leads us to prefer a slightly biased estimator to an unbiased one if the reduction in variance more than compensates the introduction of bias.

Toro-Vizcarrondo and Wallace (1968) first considered the risk of the restricted estimator (RLS) when m general linear independent hypotheses are imposed on the parameters to reduce the dimensionality of the parameter space from p to $p - m$ in the form of

$$H\beta = h . \tag{2.9}$$

The mean and variance of RLS $\tilde{\beta}$ are respectively

$$\mathcal{E}\tilde{\beta} = \beta - M(H\beta-h) \tag{2.10}$$

and

$$\text{var } \tilde{\beta} = \sigma^2\{(X'X)^{-1} - MH(X'X)^{-1}H'M\} \tag{2.11}$$

where

$$M = (X'X)^{-1}H'(H(X'X)^{-1}H')^{-1} \quad . \tag{2.12}$$

Since $\text{var}\ \hat{\beta} - \text{var}\ \tilde{\beta} = \sigma^2 MH(X'X)^{-1}H'M'$ is positive semidefinite, we can conclude that RLS always has a variance smaller than that of OLS. Let $\delta = H\beta - h$, then the bias of RLS is easily seen from (2.10) to be $-M\delta$. The risk matrix measure can be decomposed as

$$\text{MSE}(\tilde{\beta}) = \text{var}(\tilde{\beta}) + (\text{bias}\ \tilde{\beta})(\text{bias}\ \tilde{\beta})' \tag{2.13}$$

$$= \sigma^2\{(X'X)^{-1} - MH(X'X)^{-1}H'M'\} + M\delta\delta'M' \tag{2.14}$$

which is to be compared with the risk matrix measure of OLS

$$\text{MSE}(\hat{\beta}) = \text{var}(\hat{\beta}) = \sigma^2(X'X)^{-1} \quad . \tag{2.15}$$

It follows that for RLS to have a smaller risk matrix measure than OLS, a sufficient condition is

$$\lambda = \delta'[\sigma^2 H(X'X)^{-1}H']^{-1}\delta < 1 \tag{2.16}$$

that is, RLS are better than OLS if the hypothetical restrictions are nearly correct with hypothesis errors fall within the concentrated ellipsoid in (2.16).

Wallace (1972) argued that condition (2.16) is over-stringent for most econometric purposes. He observed that if within sample prediction is our objective, then the appropriate measure of goodness is the total prediction square error, $\mathcal{E}(X\tilde{\beta}-X\beta)'(X\tilde{\beta}-X\beta)$, which is the expectation of a quadratic loss defined in (1.3) with $W = X'X$. Analogous to (2.13), the decomposition of the quadratic risk $R_W(\tilde{\beta})$ is

$$R_W(\tilde{\beta}) = \text{tr}\, W\, \text{var}(\tilde{\beta}) + (\text{bias}\ \tilde{\beta})'W(\text{bias}\ \tilde{\beta}) \tag{2.17}$$

so that the total prediction square error (TPSE) for RLS and OLS are

respectively

$$R_{X'X}(\tilde{\beta}) = \sigma^2(p-m) + \delta'(H(X'X)^{-1}H')^{-1}\delta \tag{2.18}$$

and

$$R_{X'X}(\hat{\beta}) = \sigma^2 p \ . \tag{2.19}$$

Comparing (2.18) and (2.19), we find that a weaker condition for $\tilde{\beta}$ to be better than $\hat{\beta}$ is that

$$\lambda = \frac{\delta'(H(X'X)^{-1}H')^{-1}\delta}{\sigma^2} < m \ . \tag{2.20}$$

However, econometricians are quite often interested in the closeness of estimates to the true parameters, so the appropriate quadratic risk should be the unweighted one, namely, with $W = I_p$. This is the well-known total mean square error (TMSE) risk which for RLS and OLS are respectively

$$R_I(\tilde{\beta}) = \text{tr } \sigma^2\{(X'X)^{-1} - MH(X'X)^{-1}H'M\} + \delta'M'M\delta \tag{2.21}$$

and

$$R_I(\hat{\beta}) = \text{tr } \sigma^2(X'X)^{-1} \ . \tag{2.22}$$

It is seen that this criterion depends on X and the collinearity structure of the design matrix would seriously affect the relative risk of the two estimators. In general, a sufficient condition for the TMSE or $\tilde{\beta}$ to be smaller than that of $\hat{\beta}$ is

$$\lambda = \frac{\delta'(H(X'X)^{-1}H')^{-1}\delta}{\sigma^2} \leq \frac{\text{tr } MH(X'X)^{-1}H'M'}{d_L} \tag{2.23}$$

while the opposite is true if

$$\lambda \geq \frac{\text{tr } MH(X'X)^{-1}H'M'}{d_S} \tag{2.24}$$

where d_L and d_s are the largest and smallest e-root of $(X'X)^{-1}$ respectively. Unless H and $X'X$ are of special form, there is a region of indeterminacy, namely

$$\frac{\text{tr } MH(X'X)^{-1}H'M'}{d_L} < \lambda < \frac{\text{tr } MH(X'X)^{-1}H'M'}{d_s}$$

such that one cannot be sure whether RLS or OLS is superior in the TMSE sense. Note that the largest and smallest e-roots could be very far apart if columns of X are highly collinear.

§2.3 Preliminary Test Estimators

The econometrician who is not sure of the validity of the restrictions and is particularly worried about the bias of the restricted estimator when the restriction is incorrect will usually subject the hypothesis (2.9) to a statistical test. The conventional test statistic is based on the Wald principle, i.e. to test if the unrestricted estimates satisfy the hypothetical restrictions. Under the null hypothesis, the test statistic

$$F = \frac{(H\hat{\beta}-h)'(H(X'X)^{-1}H')^{-1}(H\hat{\beta}-h)}{\hat{\sigma}^2 m} \tag{2.25}$$

where

$$\hat{\sigma}^2 = (y - X\hat{\beta})'(y - X\hat{\beta})/(T - p)$$

is distributed as a *central* F-variate with m and $T - p$ degrees of freedom. Under the alternative hypothesis, F will be distributed as *non-central* F with m and $T - p$ degrees of freedom and non-centrality parameter λ defined in (2.16).

Since our object is to choose an estimator with smaller risk rather than to test whether the estimators are biased or not, we should then test if $\lambda = 1$ or $\lambda = m$ depending on our using the strong or the weak

criterion respectively rather than the conventional hypothesis of $\lambda = 0$. The upshot of this change of hypothesis is, at a given level of significance, to compare F with the critical points of non-central F distributions instead of the central F distribution. Table of critical points at .05, .10, .25 and .50 levels for the non-central F distribution with non-centrality $\lambda = 1$ is given in Wallace and Toro-Vizcarrondo (1969) for various degrees of freedom. A similar table for $\lambda = m$ is given in Goodnight and Wallace (1972). From the tables, it is noted that, for the same degrees of freedom, to test for a higher value of λ at a given level of significance is equivalent to the conventional test of $\lambda = 0$ at a lower level of significance. For instance, the 5% test of $H_o: \lambda = m$ is equivalent to a 1% test of $H_o: \lambda = 0$ when there are 1 and 18 degrees of freedom. In fact, given the parameter, there exists a family of preliminary test estimators (PTE) depending on the level of the test $(0 < \alpha < 1)$, or equivalently, the critical value of the test $(0 < c < \infty)$. One can judiciously select an appropriate critical value to obtain an "optimal" PTE. To facilitate further analysis, let us denote the PTE(c) by

$$\begin{aligned}\hat{\tilde{\beta}} &= I_{(0,c)}(F)\hat{\beta} + I_{[c,\infty)}(F)\tilde{\beta} \\ &= \hat{\beta} - I_{(0,c)}(F)M(H\hat{\beta} - h)\end{aligned} \tag{2.26}$$

where $I_{(0,c)}(F)$ and $I_{[c,\infty)}(F)$ are indicator functions which are one if F falls in the interval subscripted and zero otherwise. Bock et al (1973) obtained the mean and variance of $\hat{\tilde{\beta}}$ respectively as

$$\mathcal{E}\hat{\tilde{\beta}} = \beta - h_\lambda(2)M\delta \tag{2.27}$$

and

$$\begin{aligned}\text{var } \hat{\tilde{\beta}} = \text{var } \hat{\beta} &- \sigma^2 h_\lambda(2)MH(X'X)^{-1}H'M' \\ &+ (2h_\lambda(2) - h_\lambda^2(2) - h_\lambda(4))M\delta\delta'M'\end{aligned} \tag{2.28}$$

where

$$h_\lambda(2) = \Pr\left[\frac{\chi^2_{m+2,\lambda}}{\chi^2_{T-p}} < \frac{cm}{T-p}\right]$$

$$h_\lambda(4) = \Pr\left[\frac{\chi^2_{m+4,\lambda}}{\chi^2_{T-p}} < \frac{cm}{T-p}\right] .$$

As $\chi^2_{m+4,\lambda}$ is a non-central chi-square variable on $m + 4$ degrees of freedom, it is stochastically greater than $\chi^2_{m+2,\lambda}$, a similar variable on smaller degrees of freedom. This implies that $1 > h_\lambda(2) > h_\lambda(4) > 0$, where c is the critical value of the test. From (2.27) and (2.28), one can deduce that the matrix risk measure of PTE(c) is

$$\begin{aligned} MSE(\hat{\hat{\beta}}) = MSE(\hat{\beta}) &- \sigma^2 h_\lambda(2)MH(X'X)^{-1}H'M' \\ &+ (2h_\lambda(2) - h_\lambda(4))M\delta\delta'M' \end{aligned} \tag{2.29}$$

and that the mean prediction error of PTE(c) is

$$R_{X'X}(\hat{\hat{\beta}}) = \sigma^2[p - h_\lambda(2)m + (2h_\lambda(2) - h_\lambda(4))\lambda] . \tag{2.30}$$

Thus, according to the Wallace's weak criterion, PTE(c) is better than OLS if and only if

$$\frac{h_\lambda(2)m}{2h_\lambda(2)-h_\lambda(4)} \geq \lambda . \tag{2.31}$$

Despite the fact that the loss function is invariant and independent of X, there still exists a region of uncertainty as whether PTE(c) or OLS is superior, namely,

$$\frac{m}{2} < \lambda < m \quad \text{because} \quad 1 > \frac{h_\lambda(4)}{h_\lambda(2)} > 0 .$$

Hence, the values of λ for which PTE(c) is better than OLS are affected by the choice of the level of significance (α) or the critical value (c).

When the critical value c approaches infinity $(\alpha \to 0)$, the risk of PTE approaches that of RLS $\tilde{\beta}$. On the other hand, when c approaches 0 $(\alpha \to 1)$, the risk of PTE approaches that of OLS $\hat{\beta}$.

In the search for an optimal critical value, Sawa and Hiromatsu (1973) defined a regret function as being the difference between $R_{X'X}(\hat{\hat{\beta}})$ and $\min\{R_{X'X}(\tilde{\beta}), R_{X'X}(\hat{\beta})\}$. Observing that $R_{X'X}(\hat{\hat{\beta}})$ is unimodal they obtained the "optimal" value of c that minimise the maximum regret which turns out to be the unique fixed point solution of

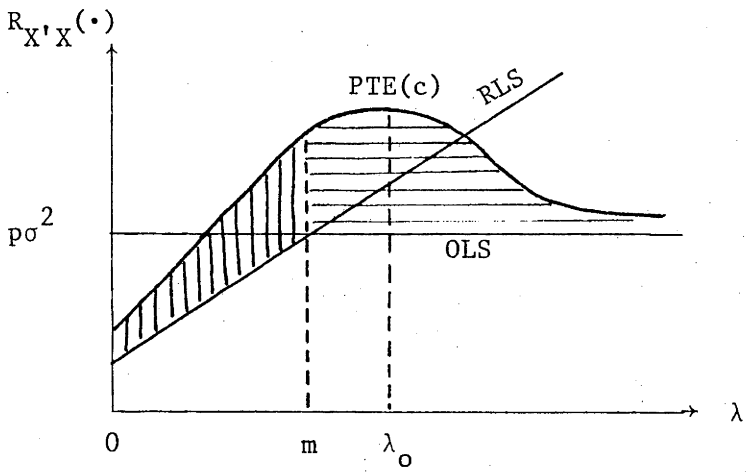
$$R_{X'X}(\hat{\hat{\beta}},c) = R^{\lambda_o}_{X'X}(\hat{\hat{\beta}},c) \tag{2.32}$$

where λ_o is the value where $R_{X'X}(\hat{\hat{\beta}})$ attains its maximum.

In the simple case of a single restriction $(m = 1)$, they found that the minimax regret critical values are about 1.88 for a wide range of degrees of freedom. For instance, at 18 degrees of freedom, the minimax critical value corresponds to the test of $\lambda = 0$ at the level of about .20 (a much higher level than the conventional ones of .05 or .10). Brook (1976) gave an alternative derivation and interpretation of the Sawa and Hiromatsu regret function and derived minimax regret optimal critical values for multiple restrictions $(m \geq 1)$ of about 2.1. It is noted that this criterion favours OLS more than RLS, which is in diametrically opposite to the weak or strong quadratic risk criteria of Wallace.

From a somewhat different point of view, Toyoda and Wallace (1974) specified a diffuse prior on λ and minimised the area bounded above by $R^{\lambda}_{X'X}(\hat{\hat{\beta}},c)$ and below by $\min\{R_{X'X}(\tilde{\beta}), R_{X'X}(\hat{\beta})\}$.

Figure 1 : Risk Functions of OLS, RLS and PTE(c)



where the shaded area is to be minimised with respect to c. The objective function can be written as

$$G(c) = \int_o^\infty [R^\lambda_{X'X}(\hat{\hat{\beta}},c) - \min\{R^\lambda_{X'X}(\tilde{\beta}), R^\lambda_{X'X}(\hat{\beta})\}]d\lambda$$

$$= \int_o^m \{\sigma^2[p-h_\lambda(2)m + (2h_\lambda(2) - h_\lambda(4))\lambda] - \sigma^2[p-m+\lambda]\}d\lambda$$

$$+ \int_m^\infty \{\sigma^2[p-h_\lambda(2)m+(2h_\lambda(2)-h_\lambda(4))\lambda] - \sigma^2 p\}d\lambda$$

$$= -m\int_o^m h^c_\lambda(2)d\lambda + \int_o^\infty \lambda\{2h^c_\lambda(2) - h^c_\lambda(4)\}d\lambda + m\int_o^m d\lambda - \int_o^m \lambda d\lambda . \quad (2.33)$$

This improper integral converges for the range $c \in [0,\infty)$ although $G(+\infty)$ does not converge. By utilizing some properties of the hypergeometric function, it can be shown that the 'optimal' or minimum average relative risk critical value is the fixed point solution of an implicit function

$$y = \phi(y)$$

where

$$y = \frac{mc}{mc+T-p}$$

$$\phi(y) = \frac{mI_y(\frac{m}{2}+1, \frac{T-p}{2}+2) + (m-4)I_y(\frac{m}{2}, \frac{T-p}{2}+1)}{(T-k+2m-2)I_y(\frac{m}{2}, \frac{T-p}{2}+1)}$$

$$I_y(\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_o^y t^{\alpha-1}(1-t)^{\beta-1}dt \ .$$

Hence, when $m \leq 4$, the "optimal" critical values are 0 implying the trivial choice of OLS. For $m > 4$, the optimal values are much smaller than the corresponding minimax regret critical values of Sawa and Hiromatsu, indicating that the minimum average relative risk criterion is more conservative than the minimax regret risk criterion. This is intuitively sensible from Figure 1. For a given value of $T - p$, the critical area to the left of m becomes larger and the area to the right becomes smaller as the values of c decreases. However, the area to the left is relatively fixed, so if m is small, then choosing small values of c is more likely to be a better decision in the absence of other information about the distribution of λ.

However, if one is to specify a prior density for λ, one might as well be completely Bayesian since the informational basis for specifying a prior distribution for the individual parameters should be better than that for the non-centrality parameter. Zellner and Vandaele (1975) postulated a priori probabilities for the null and alternative hypotheses and formulated a Bayesian minimum expected loss (BMEL) estimator by minimising the posterior expected loss

$$p_o[(\tilde{\tilde{\beta}}-h)'W(\tilde{\tilde{\beta}}-h)] + (1-p_o)[(\tilde{\tilde{\beta}}-\bar{\beta})'W(\tilde{\tilde{\beta}}-\bar{\beta}) + \&(\beta-\bar{\beta})'W(\beta-\bar{\beta})] \ .$$

with respect to $\tilde{\tilde{\beta}}$. The minimising value is

$$\tilde{\tilde{\beta}}^{*} = p_o h + (1-p_o)\bar{\beta}$$

$$= h + \left[1 - \frac{K_o}{1+K_o} \right] (\bar{\beta}-h) \qquad (2.34)$$

where p_o = posterior probability of the hypothesis H_o: $\beta = h$

W = positive definite symmetric matrix

$\bar{\beta}$ = posterior mean of β under the alternative hypothesis

H_1: $\beta \neq h$

$K_o = \frac{p_o}{1-p_o}$ is the posterior odd ratio.

Note that $\tilde{\tilde{\beta}}^{*}$ can be viewed as an estimator that shrinks the posterior mean $\bar{\beta}$ to the prior mean, h. In the usual Bayesian sense, this estimator is consistent, admissible and it minimises average risk. Leamer and Chamberlain (1976), building on the Zellner and Vandaele's result, concluded that a conditional posterior mean under the standard assumptions for the linear model is a matrix weighted average of the restricted and unrestricted estimates. This suggests the usefulness of combined estimators to which we now turn.

§2.4 Combined Estimators

The PTE $\hat{\tilde{\beta}}$ is seen to be a switching estimator between OLS $\hat{\beta}$ and RLS $\tilde{\beta}$ depending on the outcome of an essentially arbitrary test statistic. The switching function can be viewed as a weighting function for combining the two estimators which are randomly given weight zero or one. Owing to the discontinuous nature of this weighting function, Cohen (1965) showed that, under certain conditions for estimation under quadratic loss, PTE is inadmissible, although he did not suggest a superior alternative.

Huntsberger (1955) observed that PTE only utilizes the information derived from the test statistic F that it does or does not fall into the region of rejection. If more of the information contained in F is utilized, it is possible to exert more control over the uncertainty inherent in the preliminary testing procedure than is possible by merely altering the level of significance of the preliminary test. He defined the admissible class of weighting functions $\phi(F)$ as being one that satisfies:-

$$\text{(i)} \quad 0 \leq \phi(F) \leq 1 \quad \text{for all} \quad F$$

$$\text{(ii)} \quad \phi(-\sqrt{F}) = \phi(\sqrt{F}) \tag{2.35}$$

such that

$$\beta(F) = \phi(F)\hat{\beta} + (1-\phi(F))\tilde{\beta} \ .$$

If $\phi(F) = I_{[c,\infty)}(F)$, then $\beta(F)$ reduces to PTE, $\hat{\hat{\beta}}$.

It can be shown that the only unbiased weighting function is the trivial one, that is, $\phi(F) = 1$ leading to OLS. There is no uniformly minimum mean square error weighting function.

Suppose λ is fixed and we consider a scalar, $a \in [0,1]$, which can be a function of λ such that

$$\begin{aligned} \beta(a) &= a\hat{\beta} + (1-a)\tilde{\beta} \\ &= \hat{\beta} - (1-a)M(H\hat{\beta}-h) \ . \end{aligned} \tag{2.37}$$

The conditional mean and variance of $\beta(a)$ are respectively

$$\&\beta(a) = \beta - (1-a)M(H\beta-h) \tag{2.38}$$

$$\text{var } \beta(a) = \sigma^2\{(X'X)^{-1} - [2(1-a) - (1-a)^2]MH(X'X)^{-1}H'M'\} \ . \tag{2.39}$$

From (2.38) and (2.39), using (2.17), we obtain an expression of the TPSE for $\beta(a)$ as

$$R_{XX}(\beta(a)) = \sigma^2\{p - [2(1-a) - (1-a)^2]m + (1-a)^2\lambda\} \tag{2.40}$$

which can be minimised with respect to a.

Setting the derivative of (2.40) to zero gives

$$\frac{a^*}{1-a^*} = \frac{\lambda}{m} \tag{2.41}$$

where λ is defined in (2.16).

The 'optimal' combined estimator

$$\beta(a^*) = a^*\hat{\beta} + (1-a^*)\tilde{\beta}$$

$$= \tilde{\beta} + a^*(\hat{\beta} - \tilde{\beta})$$

is similar to the Zellner and Vandaele's BMEL estimator in (2.34) where a^* was interpreted as function of a priori probabilities. However, a^* is unknown as it depends on the unknown parameter λ. An *operational* combined estimator can be obtained by substituting unbiased estimators for the unknowns. By so doing, the estimates for the optimal weights is related to the test statistic F,

$$\frac{\hat{a}^*}{1-\hat{a}^*} = \frac{(H\hat{\beta}-h)'(H(X'X)^{-1}H')^{-1}(H\hat{\beta}-h)}{m\hat{\sigma}^2} = F \ .$$

Alternatively, we can write

$$\hat{a}^* = \frac{F}{1+F} = \phi^*(F) \ . \tag{2.42}$$

This is a generalization of the results of Huntsberger (1955) and Feldstein (1973) who considered the special case of single restrictions (m = 1). However, if we use $\hat{a}^*$ in (2.42) as weights for the combined

estimator (2.37), we cannot guarantee that its TPSE is minimum since the moment formulae in (2.38) and (2.39) are no longer valid when $\hat{a}^*$ is stochastic. However, simulation results for the simpler case by these two authors indicated that $\beta(\hat{a}^*)$, the operational combined estimator, is superior to OLS and PTE when λ is small and only slightly inferior for a wide range of values of λ. That means that $\beta(\hat{a}^*)$ is neither minimax nor admissible and cannot be taken as a superior alternative to OLS or PTE. In the following sections, we give an account of the most recent development of minimax estimators in the literature and discuss their usefulness in econometrics.

§2.5 Minimax Estimators Under Special Quadratic Loss

The total prediction square error (TPSE) is a special invariant quadratic loss function in the regression situation. Under such loss function, the problem of obtaining minimax estimators that is uniformly superior to the conventional OLS, $\hat{\beta}$, has been solved recently by prominent mathematical statisticians. Previous sections have been concerned with the relative superiority of the RLS and OLS which are related by

$$\tilde{\beta} = \hat{\beta} - (X'X)^{-1}H'[H(X'X)^{-1}H']^{-1}(H\hat{\beta}-h) \tag{2.43}$$

under the hypothetical restriction $H\beta = h$.

It is more convenient mathematically to consider the "reduced problem", namely, to estimate the $m\times 1$ parameter vector $\delta = H\beta - h$ from the $m\times 1$ random vector $\hat{\delta} = H\hat{\beta} - h$. Under assumption of normality of the regression errors, we have that

$$\hat{\delta} \sim N_m(\delta, \sigma^2 \Sigma) \tag{2.44}$$

where

$$\Sigma = H(X'X)^{-1}H'$$

is a known positive definite symmetric matrix. The OLS of δ is $\hat{\delta}$ while the RLS of δ is the null vector. The PTE(c) of δ is $I_{(c,\infty)}(F)\hat{\delta}$ where $F = \frac{1}{m}\hat{\delta}[\hat{\sigma}^2\Sigma]^{-1}\hat{\delta}$ is the usual test statistic for the hypothesis. Finally, the 'optimal' combined estimator is $\hat{a}^*\hat{\delta} = \frac{F}{1+F}\hat{\delta} = (1-\frac{1}{1+F})\hat{\delta}$.

The special invariant quadratic loss function of the OLS $\hat{\delta}$ for estimating δ is $(\hat{\delta}-\delta)'\Sigma^{-1}(\hat{\delta}-\delta)$ which is equivalent to the total prediction square error criterion in the β-space. As Σ is a known positive definite symmetric matrix, we can write $\Sigma = \Sigma^{\frac{1}{2}}\Sigma^{\frac{1}{2}}$. Let $\theta = \Sigma^{-\frac{1}{2}}\delta$, so that $\hat{\theta} = \Sigma^{-\frac{1}{2}}\hat{\delta}$ which has a spherical normal distribution, i.e., $\hat{\theta} \sim N_m(\theta,\sigma^2 I_m)$ while the invariant quadratic loss for $\hat{\theta}$ is simply the squared distance of the estimation error, i.e., $(\hat{\theta}-\theta)'(\hat{\theta}-\theta)$.

Under such a loss function it can be shown that OLS $\hat{\theta}$ which is minimax is no longer admissible when the dimensionality of the problem exceeds two. For $m = 1$ and 2, it is well-known that $\hat{\theta}$ is admissible. However, as m approaches ∞, it becomes less and less admissible. Let us define $|u|^2 = u'u$. Stein (1956) took notice of the fact that $\text{tr var}(\hat{\theta}) = \mathcal{E}|\hat{\theta}|^2 - |\mathcal{E}\hat{\theta}|^2 = m\sigma^2$. When $\hat{\theta}$ is unbiased, the expected length of the unbiased estimator which is $\mathcal{E}|\hat{\theta}|^2 = |\theta|^2 + m\sigma^2$ exceeds the true parameter length by an order of m, the number of dimensions of the parameter space. It is reasonable to suspect that as m increases, $\hat{\theta}$ could be a very poor estimator. Indeed, $\hat{\theta}$ is inadmissible when $m > 2$ also Stein has shown that there exist minimax alternatives. Since σ^2 has the effect of scaling the variance, we can set $\sigma^2 = 1$ without losing generality for the known variance case. The risk of OLS under the invariant loss is thus simply equal to m. James and Stein (1961) considered the following estimator

$$\theta^* = \left(1 - \frac{b}{|\hat{\theta}|^2}\right)\hat{\theta} \qquad (2.45)$$

where b is a positive constant.

It can be shown that

$$R_I(\theta^*) = \mathcal{E}|\theta^* - \theta|^2$$

$$= m + b[b-2(m-2)]\mathcal{E}\frac{1}{|\hat{\theta}|^2} . \qquad (2.46)$$

This expression is clearly smaller than m for all θ, if $m > 2$ and $0 < b < 2(m-2)$. Furthermore, if we differentiate (2.46) with respect to b, we found that $R_I(\theta^*)$ is minimised when $b = m-2$. Hence, we obtain the celebrated James-Stein estimator (JSE).

$$\theta_{JS} = \left(1 - \frac{m-2}{|\hat{\theta}|^2}\right)\hat{\theta} . \qquad (2.47)$$

The risk of θ_{JS} is 2 at $\theta = 0$ and increases monotonically with $|\theta|^2$ to the value m as $|\theta|^2 \to \infty$. Considering the fact that the usual number of parameters to be estimated easily exceeds 2, the improvement of θ_{JS} over $\hat{\theta}$ when θ is close to zero is considerable. More importantly, it can never do worse than $\hat{\theta}$ whatever the parameter value!

Although θ_{JS} is minimax, it is not admissible. This implies that there exist other estimators that have uniformly lower risk than θ_{JS}. One example is the positive part version (Stein (1966))

$$\theta_{JS}^+ = \left(1 - \frac{m-2}{|\hat{\theta}|^2}\right)^+ \hat{\theta} = I_{(m-2,\infty)}(F)\left(1 - \frac{m-2}{F}\right)\hat{\theta} \qquad (2.48)$$

where $x^+ = \max[0,x]$. With θ_{JS}^+ as a minimax alternative for the OLS $\hat{\theta}$, a simple minimax alternative for the PTE of θ is easily seen to be $I_{(c,\infty)}(F)\theta_{JS}^+$ which was proposed by Sclove et al (1972). [Here, in this special case, $F = |\hat{\theta}|^2$ has a noncentral chi-square distribution with m degrees of freedom and non-centrality parameter $|\theta|^2$]. If the critical

point of the preliminary test c equals $m - 2$, then the Sclove estimator is exactly the positive part estimator because

$$I_{(c,\infty)}(F)\theta^{+}_{JS} = I_{(c,\infty)}(F)(1 - \frac{m-2}{F})^{+}\hat{\theta} = I^{2}_{(m-2,\infty)}(F)(1 - \frac{m-2}{F})\hat{\theta} = \theta^{+}_{JS}. \quad (2.49)$$

Since $I^{2}_{(m-2,\infty)}(F) = I_{(m-2,\infty)}(F)$.

A general minimax estimator was proposed by Baranchik (1970) as

$$\theta_B = \left(1 - \frac{r(|\hat{\theta}|^2)(m-2)}{|\hat{\theta}|^2} \right) \hat{\theta} \quad (2.50)$$

where $m > 2$

$r(\cdot)$ is any nondecreasing function such that $0 \leq r(\cdot) \leq 2$

and $r(\cdot) \not\equiv 0$ or 2.

Alam (1973) determined minimax conditions $0 \leq r(|\hat{\theta}|) \leq 2$ which permit $r(\cdot)$ to decrease.

For the more general case when σ^2 is unknown, Efron and Morris (1976) proved the following result.

Theorem 2.5.1 (Minimax Estimators)

For members of the family of estimators of the form

$$\theta_{EM} = \left(1 - \frac{r(F)(m-2)}{F} \right) \hat{\theta} \quad (2.51)$$

where $F = \dfrac{|\hat{\theta}|^2}{s^2}$ $\qquad s^2 \sim \dfrac{\sigma^2}{n+2} \chi^2_n$

$r(\cdot)$ is any real valued function on $(0,\infty)$

[and absolutely continuous]

to be minimax, it is necessary and sufficient the following conditions be satisfied

(i) $0 \leq r(F) \leq 2$ for all F

(ii) $\phi_n(F) = \dfrac{F^{\ell} r(F)}{(2-r(F))^{1+2c_n}}$ is nondecreasing where $\ell = \dfrac{m-2}{2}$

$c_n = \dfrac{m-2}{n+2}$

and

(iii) if F_1 exists such that $r(F_1) = 2$, then $r(F) = 2$ for all $F > F_1$.

Remarks: these conditions are obviously weaker than Baranchik's since $r(\cdot)$ need not be nondecreasing here.

In their original proof, there are some minor mistakes which we correct and present below. To prove the theorem, we need two lemmas

LEMMA 2.5.1 (Stein's Identity)

Let $x_1 \ldots x_m$ be independent normal variates with mean θ_i $(i = 1 \ldots m)$ and common variance σ^2. Then for any absolutely continuous function $h(x)$ with Lebesgue measurable derivative $\dfrac{\partial h(x)}{\partial x}$ satisfying $\mathcal{E}\left|\dfrac{\partial h(x)}{\partial x}\right| < \infty$, we have the identity

$$\mathcal{E}(x_i-\theta_i)h(x_i) \equiv \sigma^2 \mathcal{E}\left|\frac{\partial h(x_i)}{\partial x_i}\right| \qquad i = 1 \ldots m \tag{2.52}$$

[Proof made use of integration by parts. See Stein (1973)].

LEMMA 2.5.2. (Unbiased Estimator of Risk)

Suppose $r(F)$ of the Efron and Morris estimator θ_{EM} at (2.51) is absolutely continuous with derivative $\dot{r}(F)$. If the risk of θ_{EM} is finite and if the expectation of each term in (2.53) exists, then a unique unbiased estimator of $\dfrac{1}{\sigma^2} R_I(\theta_{EM})$ based on F exists and is

$$\frac{1}{\sigma^2}\hat{R}_I(\theta_{EM}) = m-(m-2)\left[\frac{m-2}{F} r(F)(2-r(F)) + 4\dot{r}(F)(1+c_n r(F))\right] . \tag{2.53}$$

Proof. Writing $B(F) = (m-2)\frac{r(F)}{F}$ so that

$$\theta_{EM} = \hat{\theta} - B(F)\hat{\theta}$$

$$\frac{1}{\sigma^2} R_I(\theta_{EM}) = \frac{1}{\sigma^2} (\hat{\theta}-\theta-B(F)\hat{\theta})'(\hat{\theta}-\theta-B(F)\hat{\theta})$$

$$= m-2\,\mathcal{E} \sum_{i=1}^{m} \frac{B(F)\hat{\theta}_i(\hat{\theta}_i-\theta_i)}{2} + \mathcal{E}B^2(F)\,\frac{|\hat{\theta}|^2}{\sigma^2}$$

let

$$h_i(\hat{\theta}) = \hat{\theta}_i B(F)$$

where

$$F = \frac{|\hat{\theta}|^2}{s^2} = \frac{\sum_{i=1}^{m} \hat{\theta}_i^2}{s^2} .$$

Thus

$$\frac{1}{\sigma^2} R_I(\theta_{EM}) = m-2\,\mathcal{E} \sum_{i=1}^{m} \frac{h_i(\hat{\theta})(\hat{\theta}_i-\theta_i)}{\sigma^2} + \mathcal{E}B^2(F)F\,\frac{s^2}{\sigma^2}$$

and invoking Stein's identity of Lemma 2.5.1, gives

$$= m-2\,\mathcal{E} \sum_{i=1}^{m} \left| \frac{\partial h(\hat{\theta}_i)}{\partial \hat{\theta}_i} \right| + \mathcal{E}B^2(F)F\frac{s^2}{\sigma^2}$$

$$= m-2m\,\mathcal{E}B(F) - 4\,\mathcal{E}F\dot{B}(F) + \mathcal{E}B^2(F)F\,\frac{s^2}{\sigma^2} . \qquad (2.54)$$

Using the fact that

$$\mathcal{E}(W-n\psi)h(W) = 2\psi\,\mathcal{E}W\dot{h}(W) \qquad (2.55)$$

for $W \sim \psi\chi_n^2$ and $h(\cdot)$ such that the expectations in (2.55) exists. We take $W = s^2$, $\psi = \frac{\sigma^2}{n+2}$ and apply (2.55) to $h(W) = B^2(F)F$ to obtain an expression for the last term in (2.54), viz.

$$\mathcal{E}B^2(F)F\,\frac{s^2}{\sigma^2} = \frac{1}{\sigma^2}\,\mathcal{E}Wh(W)$$

$$= \frac{1}{\sigma^2}\{\mathcal{E}n\psi h(W) + 2\psi\mathcal{E}W\dot{h}(W)\} .$$

$$= \frac{n}{n+2}\mathcal{E}B^2(F)F - \frac{2}{n+2}\mathcal{E}B(F)F[B(F) + 2\dot{B}(F)F]$$

$$= \frac{n-2}{n+2}\mathcal{E}B^2(F)F - \frac{4}{n+2}\mathcal{E}F^2\dot{B}(F)B(F) . \qquad (2.56)$$

Substituting (2.56) into (2.54) and dropping expectation operator gives the unbiased estimator of $\frac{1}{\sigma^2}R_I(\theta_{EM})$

$$\frac{1}{\sigma^2}\hat{R}_I(\theta_{EM}) = m - 2mB(F) - 4F\dot{B}(F) + FB^2(F)$$
$$- \frac{4}{n+2}FB(F)[B(F) + F\dot{B}(F)] . \qquad (2.57)$$

Now, as $B(F) = (m-2)\frac{r(F)}{F}$, (2.57) can be written as

$$\frac{1}{\sigma^2}\hat{R}_I(\theta_{EM}) = m - 2m(m-2)\frac{r(F)}{F} - 4(m-2)\dot{r}(F) + 4(m-2)\frac{r(F)}{F}$$
$$+ (m-2)^2\frac{r^2(F)}{F} - \frac{4}{n+2}(m-2)^2 r(F)\dot{r}(F)$$

$$= m - (m-2)[\frac{r(F)}{F}(m-2)(2-r(F)) + 4\dot{r}(F)(1+\frac{m-2}{n+2}r(F))] \qquad (2.53)$$

Q.E.D. □

Remark: This unbiased estimator for $\frac{1}{\sigma^2}R_I(\theta_{EM})$ is unique because F is a sufficient statistic and the family of non-central F-distribution is complete.

We are now in a position to prove Theorem 2.5.1.

Proof of Theorem 2.5.1

Note that with $2\ell = m-2$, we can write

$$\frac{1}{\sigma^2}\hat{R}(\theta_{EM}) \begin{cases} = m - 4\ell F^{-\ell}(2 - r(F))^{2+2c_n}\dot{\phi}_n(F) \\ = m \quad \text{when} \quad r(F) = 2 . \end{cases} \qquad (2.58)$$

It is clear that condition (i), (ii) and (iii) are sufficient to ensure that $\frac{1}{\sigma^2}\hat{R}_I(\theta_{EM}) \leq m$ for all F and hence $R_I(\theta_{EM}) = \mathcal{E}\hat{R}_I(\theta_{EM})$ is uniformly smaller than the constant risk $\sigma^2 m$ of $\hat{\theta}$.

To prove necessity of these conditions. Suppose $\hat{R}(\theta_{EM}) \leq m\sigma^2$ for all $F > 0$. If there exists F_o such that $r(F_o) < 0$, then $\phi_n(F) < 0$, then there exists $0 < F < F_o$ such that $\dot{\phi}_n(F) < 0$ and $\hat{R}(\theta_{EM}) > m\sigma^2$. Hence, it is necessary that $r(F) \geq 0$ for all F. If there exists F_2 such that $r(F_2) > 2$ then $\dot{r}(F_2) > 0$ in order that $\hat{R}(\theta_{EM}) \leq m\sigma^2$ from (2.53). Hence, $r(F)$ increases monotonically to $r(\infty) > 2$. Thus $\lim_{|\theta|^2 \to \infty} \hat{R}(\theta_{EM}) > m\sigma^2$ implying $\hat{R}(\theta_{EM}) > m\sigma^2$ for large F when $\hat{R}(\theta_{EM})$ exists, leading to a contradiction. From (2.58) we see that condition (ii) is also necessary for $\hat{R}(\theta_{EM}) \leq m\sigma^2$ for all F. Condition (iii) must hold because $\dot{r}(F) \geq 0$ for all F such that $r(F) = 2$, from (2.53). But if $\dot{r}(F) > 0$ for some F, then there exists $F_2 > F$ such that $r(F_2) > 2$, violating condition (i).

Q.E.D. □

We see that the James-Stein estimator and its positive part version are members of the family of minimax estimators of Efron and Morris with $r(F) = 1$ and $r(F) = \min(1, \frac{F}{m-2})$ respectively, and both satisfy all the conditions. To translate these minimax estimators to the regression problem, we see that under the invariant quadratic loss function, the minimax estimator for $H\beta - h$ is

$$\delta_{EM} = \left(1 - \frac{r(F)(m-2)}{F}\right)(H\hat{\beta} - h) \tag{2.59}$$

where

$$F = \frac{(H\hat{\beta}-h)'(H(X'X)^{-1}H')^{-1}(H\hat{\beta}-h)}{s^2} = \frac{|\hat{\theta}|^2}{s^2}$$

is $\frac{m(T-p+2)}{T-p}$ times the usual likelihood ratio test statistic for the

hypothesis H_o: $H\beta = h$. If $H = I_p$ and $h = 0$, then the minimax alternative to the conventional OLS, $\hat{\beta}$ is

$$\beta_{EM} = \left(1 - \frac{r(F)(p-2)}{F} \right) \hat{\beta} \tag{2.60}$$

which is seen as an estimator which continuously and uniformly shrinks $\hat{\beta}$ to the null vector. The estimator can perform well only if h is actually the null vector. In general, if $h = \beta_o \neq 0$, then a proper minimax estimator is

$$\beta_{EM} = \beta_o + \left(1 - \frac{r(F)(p-2)}{F} \right)(\hat{\beta} - \beta_o) \tag{2.61}$$

which shrinks $\hat{\beta}$ to the a prior origin β_o.

Comparing (2.61) to the various versions of pre-test estimators and combined estimators, it is not surprising why earlier authors fail to strike at a minimax alternative to the conventional Least Squares estimator. It reveals that any ad hoc modifications of the conventional procedure could be disastrous unless the minimax conditions in Theorem 2.5.1 are met.

In summary, under invariant loss criterion, OLS could be uniformly improved upon. The above result should convince econometricians that a simple alternative like the Stein-James positive part exists and under suitable conditions, should be used as a better point estimate in place of OLS.

§2.6 Minimax Estimator Under General Quadratic Loss

For many econometric problems, a more general quadratic loss function is desirable. One quadratic loss is the total square error loss where we give the same weight to each elements of the parameters irrespective of the strength of the data. In this section, we shall consider the

general quadratic loss function $L_w = (\hat{\delta}-\delta)'W(\hat{\delta}-\delta)$ which includes the invariant quadratic loss $(W = \Sigma^{-1})$ and the total square error loss $(W = I)$ as special cases. W is a positive definite symmetric matrix.

Theorem 2.6.1

For $\hat{\delta} \sim N_m(\delta,\Sigma)$, members of the family of estimator of the form

$$\tilde{\delta} = \left(1 - \frac{cr(F)}{F} \right) \hat{\delta} \tag{2.62}$$

where $r(\cdot)$ is a real valued measurable and non-decreasing function of F

c is a positive scalar constant

are minimax under the general quadratic loss L_W if, for $m > 2$,

(i) $0 < r(F) < 2$ for all F

(ii) $$0 \leq c \leq \frac{1}{n+2} \left(\frac{\text{tr } W\Sigma}{d_L} - 2 \right) \tag{2.63}$$

(iii) $$\text{tr } W\Sigma > 2d_L \tag{2.64}$$

where $F = \dfrac{\hat{\delta}'\Sigma^{-1}\hat{\delta}}{s^2} = \dfrac{\chi^2_{m,\lambda}}{\chi^2_n}$ is proportional to the non-central F variate on m and n degrees of freedom

d_L = largest e-root of $W\Sigma$

$\lambda = \dfrac{\delta'\Sigma^{-1}\delta}{\sigma^2}$ is the non-centrality parameter

$s^2 \sim \sigma^2\chi^2_n$ is distributed independently of $\hat{\delta}'\Sigma^{-1}\hat{\delta}$ which is $\sigma^2\chi^2_{m,\lambda}$.

To prove this theorem, we need the following lemmas but as proofs of which are given in Bock (1975) we will not present them here.

LEMMA 2.6.1.

If x is distributed as $N_p(\theta,\sigma^2\Sigma)$ and $h(\cdot)$ is a real-valued measurable function from $[0,\infty)$ to $(-\infty,\infty)$, then

$$\mathcal{E}\, h\left(\frac{x'\Sigma^{-1}x}{\sigma^2}\right) x = \theta\, \mathcal{E}\,(\chi^2_{p+2,\lambda})$$

where

$$\lambda = \frac{\theta'\Sigma^{-1}\theta}{\sigma^2} .$$

LEMMA 2.6.2.

Under the same conditions as in Lemma 2.6.1 and W is a positive definite symmetric matrix

$$\mathcal{E}\, h\left(\frac{x'\Sigma x}{\sigma^2}\right) x'Wx = \sigma^2(\operatorname{tr} W\Sigma)\mathcal{E}h(\chi^2_{p+2,\lambda}) + \theta'W\theta\,\mathcal{E}h(\chi^2_{p+4,\lambda}) .$$

LEMMA 2.6.3.

Let ϕ be a real valued measurable function defined on the positive integers. Let $K \sim$ Poisson $(\frac{\lambda}{2})$, then if the expectations of both side exists,

$$\lambda\, \mathcal{E}\, \phi(K) = \mathcal{E}2K\phi(K-1) .$$

LEMMA 2.6.4.

$$\mathcal{E}h(\chi^2_n) = \mathcal{E}\,\frac{nh(\chi^2_{n+2})}{\chi^2_{n+2}} .$$

LEMMA 2.6.5.

$s : [0,\infty) \to (0,\infty)$ monotonic nondecreasing in u

$t : [0,\infty) \to (0,\infty)$ monotonic nonincreasing in u

then

$$\mathcal{E}\{s(u)[\mathcal{E}u-u]\} \le 0 \le \mathcal{E}\{t(u)(\mathcal{E}u-u)\} .$$

LEMMA 2.6.6.

$$\chi^2_{m,\lambda} \equiv \chi^2_{m+2K} \quad \text{where} \quad K \sim \text{Poisson}\ (\tfrac{\lambda}{2}) \ .$$

We are now in a position to prove Theorem 2.6.1. This proof modifies the proof of Bock's (1975) result on the more general class of spherical symmetric estimators for the special class of estimators considered by Efron and Morris so as to maintain a continuity of arguments from the preceding section.

Proof of Theorem 2.6.1

The risk of $\hat{\delta}$ is well known to be $\sigma^2 \text{tr}\, \Sigma W$. $\hat{\delta}$ is minimax because its risk is independent of the values of the true parameters. If we can show that the risk of $\tilde{\delta}$ satisfying conditions (i), (ii) and (iii) is uniformly smaller than $\sigma^2 \text{tr}\, \Sigma W$, then $\tilde{\delta}$ is minimax.

Let us write, with $B(F) = \dfrac{cr(F)}{F}$, the estimator in (2.62)

$$\tilde{\delta} = (1 - B(F))\hat{\delta} = \hat{\delta} - B(F)\hat{\delta}$$

which has risk under the quadratic loss L_W as

$$\begin{aligned} R_W(\tilde{\delta}) &= \mathcal{E}(\tilde{\delta}-\delta)'W(\tilde{\delta}-\delta) \\ &= \mathcal{E}(\hat{\delta}-B(F)\hat{\delta}-\delta)'W(\hat{\delta}-B(F)\hat{\delta}-\delta) \\ &= R_W(\hat{\delta}) + 2\mathcal{E}B(F)\hat{\delta}'W\hat{\delta} + \mathcal{E}B^2(F)\hat{\delta}'W\hat{\delta} - 2\mathcal{E}B(F)\hat{\delta}'W\hat{\delta} \ . \end{aligned} \tag{2.65}$$

Let $D_W(F) = \dfrac{1}{\sigma^2}\{R_W(\tilde{\delta}) - R_W(\hat{\delta})\}$.

If $D_W(F) \leq 0$, then $\tilde{\delta}$ has smaller risk than $\hat{\delta}$.

Conditional on s^2, $F \sim \chi^2_{m,\lambda} = \chi^2_{m+2K}$ where $K \sim P_o(\frac{\lambda}{2})$ (by Lemma 2.6.8), let us denote $B(i) = B(\chi^2_{m+2K+i})$, $i = 0,2,4$. Then conditional on s^2,

$$D_W(F) = -\frac{2}{\sigma^2}\,\mathcal{E}B(F)\hat{\delta}'W\hat{\delta} + \frac{2}{\sigma^2}\,\mathcal{E}B(F)\hat{\delta}'W\delta + \frac{1}{\sigma^2}\,\mathcal{E}B^2(F)\hat{\delta}'W\hat{\delta}$$

$$= -2\{(\mathrm{tr}\,\Sigma W)\mathcal{E}B(\chi^2_{m+2,\lambda}) + \frac{\delta'W\delta}{\sigma^2}\,\mathcal{E}B(\chi^2_{m+4,\lambda})\}$$

$$+ 2\,\frac{\delta'W\delta}{\sigma^2}\,\mathcal{E}B(\chi^2_{m+2,\lambda})$$

$$+ \{(\mathrm{tr}\,\Sigma W)\mathcal{E}B^2(\chi^2_{m+2,\lambda}) + \frac{\delta'W\delta}{\sigma^2}\,\mathcal{E}B^2(\chi^2_{m+4,\lambda})\}$$

(by Lemma 2.6.1 and Lemma 2.6.2)

$$= (\mathrm{tr}\,\Sigma W)\mathcal{E}\{B(2)[-2+B(2)] + \frac{2K\alpha(\delta)}{(\mathrm{tr}\Sigma W)}\,B(2)[-2+B(2)] + \frac{4K\alpha(\delta)}{(\mathrm{tr}\Sigma W)}\,B(0)\}$$

where

$$\alpha(\delta) = \frac{\delta'W\delta}{\sigma^2\lambda} \quad \text{and} \quad K \sim P_o(\tfrac{\lambda}{2})$$

(by Lemma 2.6.3 and Lemma 2.6.6)

$$= (\mathrm{tr}\,\Sigma W)\mathcal{E}\{(1+H)B(2)[B(2)-2] + 2HB(0)\} \qquad (2.66)$$

where $H = \dfrac{2K\alpha(\delta)}{\mathrm{tr}\,\Sigma W}$.

The *first* term inside curly bracket of (2.66) is

$$\mathcal{E}(1+H)B(2)[B(2)-2]$$

$$= \mathcal{E}\,\frac{1}{\chi^2_{m+2+2K}}\,(1+H)c\chi_n^2 r\left(\frac{\chi^2_{m+2+2K}}{\chi^2_n}\right)\left[\frac{c\chi_n^2 r\left(\dfrac{\chi^2_{m+2+2K}}{\chi^2_n}\right) - 2\chi^2_{m+2+2K}}{\chi^2_{m+2+2K}}\right]$$

$$= \mathcal{E}\,\frac{h(\chi^2_{m+2+2K})}{\chi^2_{m+2+2K}} \qquad \text{(say)}$$

where $K \sim P_o(\frac{\lambda}{2})$.

Using Lemma 2.6.4, we have

$$\mathcal{E}\frac{h(\chi^2_{m+2+2K})}{\chi^2_{m+2+2K}} = \mathcal{E}\frac{h(\chi^2_{m+2K})}{m+2K}$$

$$= \mathcal{E}\frac{1}{m+2K}(1+H)c\chi_n^2 r\left(\frac{\chi^2_{m+2K}}{\chi^2_n}\right)\frac{\left[c\chi_n^2 r\left(\frac{\chi^2_{m+2K}}{\chi^2_n}\right) - 2\chi^2_{m+2K}\right]}{\chi^2_{m+2K}}$$

$$= \mathcal{E}\frac{k(\chi^2_{m+2K})}{\chi^2_{m+2K}} \qquad \text{(say)}.$$

Applying Lemma 2.6.4 again, we have

$$\mathcal{E}\frac{k(\chi^2_{m+2K})}{\chi^2_{m+2K}} = \mathcal{E}\frac{k(\chi^2_{m+2K-2})}{m+2K-2}$$

$$= \mathcal{E}\frac{1}{(m+2K)(m+2K-2)}(1+H)c\chi_n^2 r\left(\frac{\chi^2_{m+2K-2}}{\chi^2_n}\right)\left[c\chi_n^2 r\left(\frac{\chi^2_{m+2K-2}}{\chi^2_n}\right) - 2\chi^2_{m+2K-2}\right].$$

Similarly, the *second* term inside curly bracket of (2.66) is

$$\mathcal{E}2HB(0) = 2Hc\mathcal{E}\frac{\chi_n^2}{\chi^2_{m+2K}} r\left(\frac{\chi^2_{m+2K}}{\chi^2_n}\right)$$

$$= 2Hc\mathcal{E}\frac{\chi_n^2}{m+2K-2} r\left(\frac{\chi^2_{m+2K-2}}{\chi^2_n}\right) \quad \text{(by Lemma 2.6.4)}.$$

Thus, combining these two terms, we have

$$D_W(F) = (\mathrm{tr}\Sigma W)\mathcal{E}\frac{c\chi_n^2}{(m+2K)(m+2K-2)} r\left(\frac{\chi^2_{m+2K-2}}{\chi^2_n}\right)\left\{\left[c\chi_n^2 r\left(\frac{\chi^2_{m+2K-2}}{\chi^2_n}\right) - 2\chi^2_{m+2K-2}\right](1+H)\right.$$

$$\left. + 2H(m+2K)\right\}.$$

If $0 < r(\cdot) < 2$, then

$$D_W(F) \leq (\mathrm{tr}\Sigma W)\,\mathcal{E}\,\frac{c\chi_n^2}{(m+2K)(m+2K-2)}\, r(\cdot)\{[c\chi_n^2 - \chi^2_{m+2K-2}]2(1+H) + 2H(m+2K)\}\ .$$

Now, conditional on χ^2_{m+2K-2}, the upper bound of $D_W(F)$ is seen to be a real valued function of χ_n^2. Applying Lemma 2.6.4, we have

$$\begin{aligned} D_W(F) &\leq n\,c\,\mathcal{E}\,\frac{\mathrm{tr}(\Sigma W)}{(m+2K)(m+2K-2)}\, r(\cdot)\{[m-2+2K-\chi^2_{m+2K-2} + c(\chi^2_{n+2}-n-2)]2(1+H) \\ &\qquad + 2(1+H)c(n+2) - 2(m-2+2K) + 4H\} \\ &= n\,c\,\mathcal{E}\,\frac{\mathrm{tr}(\Sigma W)}{(m+2K)(m+2K-2)}\, r(\cdot)\{[m-2+2K-\chi^2_{m-2+2K} + c(\chi^2_{n+2}-n-2)]2(1+H) \\ &\qquad + 2c(n+2) - 2(m-2) + 2H[c(n+2) + 2 - \tfrac{2K}{H}]\}\ . \end{aligned}$$

As $r\left(\dfrac{\chi^2_{m+2K-2}}{\chi^2_{n+2}}\right)$ is known to be a monotonic nondecreasing function of χ^2_{m+2K-2} but a monotonic nonincreasing function χ^2_{n+2}, then we can apply Lemma 2.6.7 and deduce that

$$\mathcal{E}\{r(\cdot)[m-2+2K-\chi^2_{m-2+2K}]\} \leq 0 \leq c\,\mathcal{E}\{r(\cdot)[(n+2)-\chi^2_{n+2}]\}$$

$$D_W(F) \leq n\,c\,\mathcal{E}\,\frac{2\,\mathrm{tr}(\Sigma W)}{(m+2K)(m+2K-2)}\, r(\cdot)\{c(n+2)-(m-2)+H[c(n+2) - (\frac{\mathrm{tr}\,\Sigma W}{\alpha(\delta)} - 2)]\}. \quad (2.67)$$

Since $\alpha(\delta) = \dfrac{\delta' W \delta}{\sigma^2 \lambda} = \dfrac{\delta' W \delta}{\delta' \Sigma^{-1} \delta}$

$$= \frac{z' \Sigma^{\frac{1}{2}} W \Sigma^{\frac{1}{2}} z}{z'z} \leq d_L$$

where $z = \Sigma^{-\frac{1}{2}}\delta$

d_L = largest e-root of $W\Sigma$.

The right side of the inequality will be negative (for $m > 2$) iff

$$0 < c < \frac{1}{n+2}\left(\frac{tr\,\Sigma W}{d_L} - 2\right) \tag{2.68}$$

and

$$\frac{tr\,\Sigma W}{d_L} > 2 \tag{2.69}$$

Q.E.D. □

The case when $W = I$ was established by Bock (1975) and, when $\Sigma = W = I$ and σ^2 is known was established by Baranchik (1970) and Efron and Morris (1973). However, when σ^2 is unknown, the conditions (2.63) and (2.64) can be reformulated for the case $\Sigma = W = I$ as

(ii)* $0 < c < \frac{m-2}{n+2}$ and (iii)* $m > 2$ respectively.

When these conditions are translated back to the regression problem, we found that $\hat{\beta} \sim N_p(\beta, \sigma^2(X'X)^{-1})$ so that for the simple mean square error criterion, $W = I$, d_L becomes the largest e-root of $(X'X)^{-1}$. Then for

$$\tilde{\beta} = \left(1 - \frac{cr(F)}{F}\right)\hat{\beta}$$

to be a minimax alternative to $\hat{\beta}$, it is necessary and sufficient that for $m = p > 2$ the following conditions are satisfied.

(i) $0 < r(F) < 2$ for all F

(ii) $0 < c < \frac{1}{T-p+2}\left(\frac{tr(X'X)^{-1}}{d_L} - 2\right)$

and

(iii) $tr(X'X)^{-1} > 2d_L$.

If the data is highly collinear, then $X'X$ will be near singular and d_L would generally be very large making it difficult to satisfy

conditions (ii) and (iii) above.

The crude Stein's estimator at (2.45) with $r(\cdot) \equiv 1$ and $c = bp/(T-p)$ so that in order to satisfy the conditions we must restrict the value of b in the range

$$0 < b < \frac{(T-p)}{p(T-p+2)} \left(\frac{tr(X'X)^{-1}}{d_L} - 2 \right) .$$

If $X'X = I$, the optimal value of b was found to be $\frac{(T-p)(p-2)}{p(T-p+2)}$ which is always less than 1. If X is multicollinear, the optimal b will be very close to zero, so that $\tilde{\beta}$ is close to $\hat{\beta}$. Note that these minimax estimators are not admissible as can be seen in the next section.

§2.7 Admissible Minimax Estimators

In order to make sure that we have the "best" possible minimax alternative to the usual estimator, $\hat{\delta}$, we must aim at a minimax estimator that is also admissible. Stein (1956) showed that if a spherically symmetric estimator is admissible as compared with all other spherically symmetric estimators, then it is admissible (in the class of all estimators). If the dimensionality (m) of the parameter space is less than two, the usual invariant estimator is admissible. But for $m > 2$, it was shown that there exist spherically symmetric estimators which are strictly better than $\hat{\delta}$ indicating that the latter is inadmissible in higher dimension problems. However, these minimax substitutes for $\hat{\delta}$ are not admissible. Strawderman and Cohen (1971) building on Brown's (1971) work has been able to determine a necessary and sufficient condition for an estimator having bounded risk to be admissible.

Consider again the m-mean problem of the model ($\sigma^2 = 1$ case)

$$x \sim N_m(\theta, I_m) .$$

The usual estimator for θ is $\hat{\delta} = x$ with risk equals to m.

Let G be any non-negative Borel measure on E^m, the m-dimensional Euclidean space. If, in addition, G is a finite measure, define the integrated risk of an estimator $\tilde{\delta}$ by

$$B_G(\tilde{\delta}) = \int R(\tilde{\delta},\theta)G(d\theta) \quad . \tag{2.70}$$

[If $G(E^m) = 1$, $B_G(\tilde{\delta})$ is the Bayes risk of $\tilde{\delta}$ with respect to G] whether or not G is finite one can define the generalised Bayes estimator δ_G by

$$\delta_G = \frac{\int \theta f(x|\theta)G(d\theta)}{\int f(x|\theta)G(d\theta)} \tag{2.71}$$

as long as the integrals in the above expression exist.

Let us consider estimation problem in spherically symmetric multivariate case ($m \geq 1$), the loss is the sum of square errors (it is the same as the invariant quadratic loss in this case).

Theorem 2.7.1 (Strawderman and Cohen (1971))

An estimator of the form $\delta(x) = h(|x|^2)x$ is generalised Bayes if and only if there exists a measure $F(\theta)$ such that

$$\exp\left[\frac{1}{2}\int_0^{|x|^2}(h(y)-1)dy\right] = \int \exp[-\tfrac{1}{2}|x-\theta|^2]dF(\theta) \tag{2.72}$$

(Remark: this ensures that $\delta(x)$ will have the form of a generalised Bayes estimator).

Proof. If $\delta(x) = h(|x|^2)x$ is generalised Bayes with respect to the prior $F^*(\theta)$, then from (2.71), $\delta(x)-x = \frac{d \ln g(|x|^2)}{dx}$,

$$\delta(x) = h(|x|^2)x$$

$$= \left[\frac{2\dot{g}(|x|^2)}{g(|x|^2)} + 1 \right] x \qquad (2.73)$$

where

$$g(|x|^2) = \int e^{-\frac{1}{2}|x-\theta|^2} \, dF^*(\theta)$$

$$\dot{g}(|x|^2) = \int \theta e^{-\frac{1}{2}|x-\theta|^2} \, dF^*(\theta)$$

$$\frac{d \, \ell n \, g(|x|^2)}{dx_i} = \frac{2\dot{g}(|x|^2)}{g(|x|^2)} x_i \qquad i = 1 \ldots m \, .$$

We have, from (2.73)

$$\tfrac{1}{2}[h(|x|^2)-1] = \frac{\dot{g}(|x|^2)}{g(|x|^2)}$$

$$\tfrac{1}{2}[h(y)-1] = \frac{\dot{g}(y)}{g(y)} = \frac{d}{dy} \log g(y) \, . \qquad (2.74)$$

The continuity of the partial derivatives of $g(|x|^2)$ implies that $\frac{d}{dy} g(y)$ is continuous, we integrate (2.74) to obtain

$$g(|x|^2) = k e^{\frac{1}{2}\int_o^{|x|^2} (h(y)-1)dy} = \int_{E^m} e^{-\frac{1}{2}|x-\theta|^2} \, dF^*(\theta) \, .$$

Absorbing the integrating constant $\frac{1}{k}$ into the measure $F^*(\theta)$ we obtain

$$\exp\left[\frac{1}{2} \int_o^{|x|^2} (h(y)-1)dy \right] = \int_{E^m} e^{-\frac{1}{2}|x-\theta|^2} dF(\theta) \, .$$

This proves the necessity portion of the theorem.

To prove sufficiency, assume that

$$\exp\left[\frac{1}{2} \int_o^{|x|^2} (h(y)-1)dy \right] = \int_{E^m} e^{-\frac{1}{2}|x-\theta|^2} \, dF(\theta)$$

for some measure $F(\theta)$. Then, taking logarithms, we find

$$\frac{1}{2}\int_0^{|x|^2}(h(y)-1)dy = \log\int_{E^m} e^{-\frac{1}{2}|x-\theta|^2}dF(\theta) \ .$$

Partially differenting both sides under the integral sign, we have for almost all x

$$h(|x|^2)x_i = \frac{\int_{E^m}\theta_i e^{-\frac{1}{2}|x-\theta|^2}dF(\theta)}{\int_{E^m} e^{-\frac{1}{2}|x-\theta|^2}dF(\theta)} \qquad i = 1 \ldots m \qquad (2.75)$$

This is the i'th element of $\tilde{\delta}(x)$, which is now in the form of a generalised Bayes estimator with respect to $F(\theta)$.

Q.E.D. □

Theorem 2.7.2 (Strawderman and Cohen (1971))

Necessary and sufficient conditions for a bounded risk spherically symmetric estimator $\tilde{\delta}(x) = h(|x|^2)x$ to be admissible are

(i) $$g(|x|^2) = \exp\left[\frac{1}{2}\int_0^{|x|^2}(h(y)-1)dy\right] \equiv \int_{E^m} e^{-\frac{1}{2}|x-\theta|^2}dF(\theta)$$

for some measure $F(\theta)$ and (i.e. it is generalised Bayes)

(ii) $$\int^{\infty}\frac{1}{r^{m-1}}g(r^2)dr = \infty$$ (i.e. its risk function is bounded).

Proof. We note that a (proper) Bayes estimator for θ with respect to a (proper) prior distribution is unique and hence admissible.

It is not true, however, that a unique generalised Bayes estimator must be admissible. One startling counter example was produced by Stein (1956) who showed that the usual estimator, which is generalised Bayes

with respect to the Lebesgue measure on E^m, is inadmissible when $m \geq 3$. However, results of Sacks (1963) and Brown (1971) indicates that any admissible estimator must be generalised Bayes and that a bounded risk generalised Bayes estimator with respect to a general prior distribution $F(\theta)$ is admissible iff the conditions stated in Theorem 2.7.2. are satisfied.

Condition (i) implies that $\tilde{\delta}(x)$ is generalised Bayes with respect to the prior $F(\theta)$ and condition (ii) is necessary and sufficient for a bounded risk, generalised Bayes estimator to be admissible. Condition (i) is also necessary for admissibility since if it fails $\delta(x)$ cannot be generalised Bayes, and hence not admissible.

Q.E.D. □

For example, the estimator considered by Stein (1956) and James and Stein (1961),

$$\delta(x) = \left(1 - \frac{m-2}{|x|^2} \right) x$$

which has uniformly the lowest risk among all estimator of the form

$$\delta(x) = \left(1 - \frac{b}{|x|^2} \right) x \qquad b \geq 0$$

is inadmissible. Since, at the origin the estimator is badly discontinuous and the discontinuity cannot be removed by any redefinition of the estimator on a set of measure zero. So these estimators are not generalised Bayes and consequently inadmissible.

Strawderman (1971) showed that there are no *proper* Bayes estimator for dimension less than 5, hence there is no minimax alternative to OLS that is admissible. He also demonstrated the existence when $m \geq 5$ of *proper* Bayes (and therefore admissible) estimators of θ which are also

members of the class of minimax estimators of Baranchik (1970). However, Faith (1978) extended Strawderman's result showing that admissible generalised Bayes estimators do exist for $m = 3$ and 4. His result is simplified below but some of his notations are adopted to facilitate comparison.

Let the unknown parameter θ have a multivariate normal conditional prior density $N_m(0, \frac{1-\mu}{\mu} I)$. Further, μ itself has marginal prior density $g(\mu)$, $1 \geq \mu \geq 0$, which is permitted to have an infinite integral. If $g(\mu)$ is not proper then the marginal prior density of θ

$$\pi(\theta) = \int_0^1 \pi(\theta|\mu)g(\mu)d\mu \qquad (2.77)$$

will not be proper either, by Fubini's theorem. On the other hand, if $g(\mu)$ is such that the marginal density of the sample $f(x)$ is finite for all x, then the posterior density of θ,

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{f(x)}$$

is a proper density, since the denominator of this is equal to the integral of the numerator with respect to θ. It can then be assumed that either $g(\mu)$ is proper or, if it is improper, that $g(\mu)$ is sufficiently well behaved that $\pi(\theta|x)$ is proper.

LEMMA 2.7.1.

Under the above specification of the marginal prior density of θ

$$\pi(\theta) = \int_0^1 \pi(\theta|\mu)g(\mu)d\mu \ ,$$

the Bayes estimator of θ is

$$\theta^* = \left(1 - \frac{r(|x|^2)(m-2)}{|x|^2} \right) x \qquad (2.78)$$

where

$$r(|x|^2) = \frac{|x|^2}{m-2}\, \mathcal{E}(\mu|x) \ . \tag{2.79}$$

Proof.[1] $\theta^* = \mathcal{E}(\theta|x) = [\mathcal{E}(\theta|\mu,x)|x]$

$= \mathcal{E}[(1-\mu)x|x]$

$= \mathcal{E}[1-\mu|x]x$

$= (1 - \mathcal{E}[\mu|x])x$

Q.E.D. □

Theorem 2.7.4

Let the marginal prior density $g(\mu)$ of μ satisfy $\lim_{\mu\to 1} g(\mu) < \infty$. If $g(\mu)$ is differentiable in μ and if

$$h(\mu) = \frac{\dot{g}(\mu)}{g(\mu)}\, \mu \tag{2.80}$$

is well defined for $1 > \mu > 0$, then the Bayes estimator $\theta^* = \mathcal{E}(\theta|x)$ has risk uniformly smaller than m (or minimax) if $h(\mu)$ satisfies

$$\begin{aligned} &\text{(i)} \quad h(\mu) \text{ is nonincreasing in } \mu \text{ and} \\ &\text{(ii)} \quad \lim_{\mu\to 0} h(\mu) \leq \frac{m}{2} - 3 \ , \quad 1 > \mu > 0 \ . \end{aligned} \tag{2.81}$$

---

[1] $\because$ $\mathcal{E}(\theta|\mu,x)$ is the posterior mean vector of $x \sim N(\theta, I)$

$$\theta \sim N(0, \frac{1-\mu}{\mu} I)$$

$= (I + \frac{\mu}{1-\mu} I)^{-1}(x + \frac{\mu}{1-\mu} 0)$

$= (1-\mu)x$.

Proof. To prove θ^* is minimax, it suffices to establish that $r(\cdot)$ is a real valued nondecreasing function such that $0 < r(\cdot) < 2$. [See Baranchik (1970)].

$$r(|x|^2) = \frac{|x|^2}{m-2}\, \mathcal{E}[\mu|x]$$

$$= \frac{|x|^2}{m-2}\, \frac{\int_0^1 \mu g(\mu) f(x|\mu)\, d\mu}{\int_0^1 g(\mu) f(x|\mu)\, d\mu} .$$

Since $(x-\theta)|_\mu \sim N_m(0,I)$ and $\theta|_\mu \sim N_m(0, \frac{1-\mu}{\mu} I)$ are independent, given μ, it follows that

$$x|_\mu = (x-\theta) + \theta|_\mu \sim N_m(0, \frac{1}{\mu} I) .$$

Therefore,

$$f(x|\mu) = \frac{1}{(2\pi)^{m/2}}\, \mu^{m/2} e^{-\frac{1}{2}\mu|x|^2} . \tag{2.82}$$

After substitution, we have

$$r(|x|^2) = \frac{|x|^2}{m-2} \cdot \frac{\int_0^1 \mu^{\frac{m}{2}+1} g(\mu) e^{-\frac{1}{2}\mu|x|^2}\, d\mu}{\int_0^1 \mu^{\frac{m}{2}} g(\mu) e^{-\frac{1}{2}\mu|x|^2}\, d\mu} .$$

Integrate the numerator by parts to obtain

$$r(|x|^2) = \frac{2}{m-2}\left[(\tfrac{m}{2}+1) + \frac{\int_0^1 \mu^{\frac{m}{2}+1} \cdot \dot{g}(\mu) e^{-\frac{1}{2}\mu|x|^2}\, d\mu}{\int_0^1 \mu^{\frac{m}{2}} g(\mu) e^{-\frac{1}{2}\mu|x|^2}\, d\mu} - \frac{g(1) e^{-\frac{1}{2}|x|^2}}{\int_0^1 \mu^{\frac{m}{2}} g(\mu) e^{-\frac{1}{2}\mu|x|^2}\, d\mu} \right]$$

$$= \frac{2}{m-2}\left[\left(\frac{m}{2}+1\right) + \mathcal{E}[h(\mu)|x] - \frac{g(1)}{\int_0^1 \mu^{\frac{m}{2}} g(\mu)\exp[\tfrac{1}{2}(1-\mu)|x|^2]d\mu}\right]. \quad (2.83)$$

This function is nondecreasing in $|x|^2$ because the first term is a constant and the third term is either zero (if $g(1) = 0$) or an increasing function of $|x|^2$ and hence non-negative. The second term is nondecreasing in $|x|^2$ because this equals $\mathcal{E}[h(\mu)|x]$, and $f(\mu|x)$ has monotone likelihood ratio with respect to the parameter $-|x|^2$.

Since $r(|x|^2) = \frac{|x|^2}{m-2}\,\mathcal{E}[\mu|x]$ and $Pr[\mu > 0] = 1$, we have under assumption (i) $r(|x|^2) > 0$, $h(\mu)$ is nonincreasing in μ, it follows from (2.83) that

$$r(|x|^2) \leq \frac{2}{m-2}\left[\left(\frac{m}{2}+1\right) + h(0^+)\right]$$

$$\leq 2 . \qquad \text{(by assumption (ii))}$$

Finally, $r(|x|^2) \not\equiv 2$. If $r(|x|^2) \equiv 2$, then it by (2.83) must be true both that $g(1) = 0$ and that $\mathcal{E}[h(\mu)|x] = \frac{m}{2} - 3$. This implies that

$$\lim_{\mu\to 1} g(\mu) = 0$$

and

$$h(\mu) = \frac{\dot{g}(\mu)}{g(\mu)}\,\mu = \frac{m}{2} - 3 .$$

These conditions are mutually exclusive since the second one implies that

$$g(\mu) = k\mu^{\frac{m}{2}-3} \qquad \text{where} \qquad \lim_{\mu\to 1} g(\mu) = k > 0 .$$

That shows that the estimator θ^* in (2.78) satisfies all conditions in Theorem 2.5.1 and hence θ^* is minimax.

Q.E.D. □

A special case of $g(\mu)$ is considered by Strawderman (1971) where

$$g(\mu) = a\mu^{a-1} \quad 0 \leq \mu \leq 1 \quad 0 < a < \begin{cases} 1 & \text{if } m \geq 6 \\ \frac{1}{2} & \text{if } m \geq 5 \end{cases}$$

Strawderman (1972) showed that when $m < 5$ there exist no proper Bayes minimax estimators of θ. However, the above result of Faith (1978) shows that there are improper priors for which the Bayes rule is minimax.

In summary, we can uniformly improve upon the usual estimator $\hat{\delta}$ in the normal regression problem under general quadratic loss functions. To search for admissible minimax estimators, we must confine ourselves to the class of generalised Bayes rules.

However, one interesting result of Strawderman and Cohen (1971) is that all bounded-risk generalised Bayes estimator of the form

$$\delta(x) = h(|x|^2)x$$

is admissible if it is a 'shrinker' and inadmissible if it is an 'expander' which are defined respectively as $\delta(x)$ with

$$0 \leq \limsup_{|x|\to\infty} h(|x|^2) < 1$$

and

$$\liminf_{|x|\to\infty} h(|x|^2) > 1 \ .$$

The estimator $\delta(x) = \left[1 - \frac{m}{|x|^2}\right] x$ can be shown as the dividing line between proper Bayes and improper Bayes estimator. Generalised Bayes estimators which "shrink" this estimator sufficiently for large $|x|$ are proper Bayes and hence admissible and those which "expand" this estimator are not proper Bayes.

Note that shrinkage towards zero was implied in the above estimators. In general, the results can be extended to estimators of the form

$$\delta(x) = \theta_o + h(|x-\theta_o|^2)(x-\theta_o)$$

where θ_o is an arbitrary origin in E^m.

We shall see in the next chapter that the mythical ridge estimator which appears to perform well in certain conditions has the property of a generalised Bayes "shrinker" - unfortunately, of unbounded risk and so is admissible but not minimax.

CHAPTER 3

MULTICOLLINEARITY AND RELATED ESTIMATION TECHNIQUES

§3.1 Perfect Collinearity and the Problem of Identification

Perfect collinearity is said to exist in a multiple regression model when the $(T\times p)$ matrix X has a rank less than the number of columns p. A subset of the columns is linearly dependent causing the $(p\times p)$ matrix $X'X$ in the normal equation to be singular and that the OLS estimator $\hat{\beta}$ not to exist. Suppose that the linear dependence of the columns of X is represented by

$$Xc = 0 \ , \tag{3.1}$$

where c is a non-null $p\times 1$ vector. This implies that the $T\times(p+1)$ observation matrix $[y \ X]$ satisfies a second linear relation besides the basic equation $y = X\beta + \varepsilon$. Thus, the two specifications $\&y = X\beta$ and $\&y = X\beta^*$, for $\beta \neq \beta^*$, are observationally equivalent[1] when $X\beta \equiv X\beta^*$, this happens if $\beta^* = \beta + kc$ (k being a nonvanishing scalar constant). In the circumstances, we can only obtain unique unbiased estimates for certain linear combinations of the elements of β. Such linear combinations are known as "estimable" functions. Consider that a particular linear combination, $w'\beta$ where w is a given p-element vector, is to be estimated by a linear function $a'y$. It is well-known that a necessary and sufficient condition for the estimator to be unbiased for $w'\beta$ is

$$a'X = w' \tag{3.2}$$

---

[1] Observational equivalence is interpreted as such that two different parameters give rise to the same likelihood function on the same data set.

This means that w must lie in the row-space of X. Note that this is always the case when X has full column rank because a can then be replaced by $X(X'X)^{-1}w$. In general, if (3.2) holds, the unique minimum variance linear unbiased estimator of $w'\beta$ is $w'z$, where z is any solution of the normal equations $X'Xz = X'y$ [see Rao (1945)]. This is essentially a mathematical result indicating the lack of information from the data about linear functions whose weights do not satisfy (3.2). Indeed there is a formal connection between "estimability" of linear function of parameters in single equation model and "identifiability" of linear function of structural parameters in the simultaneous econometric model. Haavelmo (1950) and Chipman (1964) related the problem of multicollinearity to the existence of other underlying structural relationships among the explanatory variables which form the basis of modern theory of 'identification' in simultaneous equations methods.

To see the connection, let X have rank $n < p$. Without loss of generality, we partition $X = [Y \quad Z]$ so that Z is $T\times n$ of full column rank and Y is $T\times m$ $(m+n = p)$. The deficiency in rank of X is m, which can be represented by a system of m independent homogenous equations

$$XA = 0 \tag{3.3}$$

where A is $(p\times m)$ of rank m.

Partitioning the p rows of A conformably into m and n rows, $A = \begin{bmatrix} B \\ \Gamma \end{bmatrix}$, (3.3) can then be written as $YB + Z\Gamma = 0$, where B is $(m\times m)$ nonsingular. Let $\Pi = -\Gamma B^{-1}$, so that $Y = Z\Pi$ expresses Y in terms of the column vectors of Z. Now, consider the related simultaneous equation model with one behavioural equation and m identities,

$$\underset{T\times 1}{y} = \underset{T\times p\times 1}{X\alpha} + \varepsilon = Y\beta + Z\gamma + \varepsilon \quad \text{where} \quad \underset{p\times 1}{\alpha} = \begin{pmatrix} \beta \\ \gamma \end{pmatrix}$$

$$0 = \underset{T\times p\times m}{XA} = YB + Z\Gamma \ . \tag{3.4}$$

The full simultaneous equation system can be written

$$[y \quad X] \begin{bmatrix} 1 & 0 \\ -\alpha & -A \\ p\times 1 & p\times m \end{bmatrix} = [y \quad \underset{1}{Y} \vdots \underset{m}{Z}] \begin{bmatrix} 1 & 0 \\ -\beta & -B \\ m\times 1 & m\times m \\ \cdots & \cdots \\ -\gamma & -\Gamma \\ n\times 1 & n\times m \end{bmatrix} = [\varepsilon \quad 0] \ . \tag{3.4'}$$

The columns of $[y \quad Y]$ can be interpreted as observations on the $(m+1)$ jointly dependent or endogenous variables, and the columns of Z as observations on the predetermined or exogenous variables. The equation $y = X\alpha + \varepsilon$ is then the "first" equation of a system of $m+1$ equations. In order that the structural parameters in this "first" equation to be identifiable, a sufficient condition [see Johnston (1972)] is that there should exist a set of m independent linear restrictions $H\alpha = h$ which can also be written as

$$(h \quad H) \begin{pmatrix} 1 \\ -\alpha \end{pmatrix} = 0$$

so that the corresponding rank criterion matrix

$$(h \quad H) \begin{bmatrix} 1 & 0 \\ -\alpha & A \end{bmatrix} = [0 \quad -HA] \tag{3.5}$$

has rank m. This is equivalent to requiring that the $(m\times p)$ matrix H has full row rank. Note that H must not be in the column space of X, otherwise HA would vanish, due to (3.3). Now, the relation between the concept of "estimable" function and the concept of "identifiable"

function can be shown. Since B is non-singular, one can use the second relation of (3.4) to get $Y = -Z\Gamma B^{-1} = Z\Pi$ and that the reduced form of the "first" equation becomes

$$y = Z(\Pi\beta+\gamma) + \varepsilon = Z\pi + \varepsilon \tag{3.6}$$

where

$$\pi = \Pi\beta + \gamma = -\Gamma B^{-1}\beta + \gamma \ .$$

Since π is $m\times 1$, incorporating the identities into the 'first' structural equation is equivalent to restricting the p-dimensional parameter space for α to the m-dimensional space. The relationship between α and the reduced form parameters can be explicitly stated as

$$\underset{p\times 1}{\alpha} = \begin{pmatrix} \underset{m\times 1}{\beta} \\ \underset{n\times 1}{\gamma} \end{pmatrix} = \begin{pmatrix} \beta \\ \pi-\Pi\beta \end{pmatrix} = \begin{pmatrix} 0 & I_m \\ -\pi & -\Pi \end{pmatrix} \begin{pmatrix} \underset{1\times 1}{-1} \\ \underset{m\times 1}{\beta} \end{pmatrix} .$$

It is seen that a given linear combination of the coefficients of the 'first' structural equation, say $w'\alpha$, can be written as

$$w'\alpha = [w_m' \quad w_n'] \begin{pmatrix} 0 & I_m \\ -\pi & -\Pi \end{pmatrix} \begin{pmatrix} -1 \\ \beta \end{pmatrix} = [-w_n'\pi \quad w_m'-w_n'\Pi] \begin{pmatrix} -1 \\ \beta \end{pmatrix} .$$

This linear function is said to be identified if it can be uniquely determined by some given values of π and Π, independently of β. A necessary and sufficient condition is that

$$w_m' - w_n'\Pi = 0 \ . \tag{3.7}$$

This is equivalent to requiring that

$$w' \begin{bmatrix} I_m \\ -\Pi \end{bmatrix} = w' \begin{bmatrix} I_m \\ \Gamma B^{-1} \end{bmatrix} = w'AB^{-1} = 0 \ . \tag{3.8}$$

Thus, w must be orthogonal to column vectors of A or that w' must be in the row space of X, i.e. $w' = a'X$ for some $a \neq 0$. This is obviously the same as (3.2), the condition for $w'\alpha$ to be "estimable".

Note that we have used the term "estimable" loosely to mean that unique least square estimates can be obtained from the 'first' equation, which of course would not be unbiased nor consistent if the structure is embedded in an interdependent system. Other treatments on the subject of identifiability such as Rothenberg (1971) and Richmond (1974) did not relate their results to the problem of multicollinearity because they are not concerned with the estimation of the equation directly by least squares. However, they have assumed that sufficient prior information in the form of constraints on the structural parameters are available so that the structural parameters can be uniquely determined by a given set of estimates on the reduced form parameters. If the exogenous variables are collinear, then some linear combination of the reduced form parameters will not be estimable and consequently the structural parameters (even if identified) cannot be estimated either. This agrees with Rothenberg's result that the reduced form parameters are themselves not identifiable because their information matrix is singular when the exogenous variables are perfectly collinear.

Farrar and Glauber (1967) mentioned the paradox that the econometrician is more fortunate if he has a problem of perfect collinearity than one whose data are almost so. The reason is that perfect collinearity can be immediately apprehended from a matrix singularity return of a computer run, which warns him of the limitation of the least square procedure in this situation. Whereas in the latter's problem, there is nothing to prevent the application of the least square procedure. Difficulty in interpreting estimated results is an almost certainly if one ignores the collinearity problem.

There is no solution to the perfect collinearity problem without additional data or prior information since the sample data are simply not "designed" to provide enough information. However, the failure to obtain LS estimates saves us from making misleading conclusions from unreliable estimates as one would have done in the case of "harmful" collinearity which is to be discussed next. This also signals serious misspecification of the data generating process.

§3.2 Harmful Collinearity and Their Solutions[2]

Collinearity is said to be 'harmful' when the classical LS estimates are highly unreliable leading to very wide confidence intervals for individual coefficients. Especially, if an econometrician tries to rely on a series of preliminary t-tests for model building, he is more likely to be led to an underspecified econometric model [see Liu (1960)]. The consequences are biased estimates, poor predictions outside the sample period and conflicting inference made by different econometricians. If LS estimation is to be a useful empirical tool, necessary remedial action must be taken in such a circumstance. The concept of "harmful multicollinearity" is difficult to define precisely, as it depends subjectively on the degree of accuracy required from the analysis and the purpose of particular problems in question.

To see the "harm" done by multicollinearity we consider first the model with two regressors ($p = 2$). The LS estimates are provided by a solution of the normal equations

$$\begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} m_{y1} \\ m_{y2} \end{pmatrix} \tag{3.9}$$

---

[2] Part of this section is reported in Lee (1978b).

where m_{ij} is the sample cross product between regressors X_i and X_j (measured as deviates from their respective sample mean)

m_{yi} is the sample cross product between y and the regressor X_i $(i = 1,2)$.

Define

$$C = \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix}^{-1} = \frac{1}{1-r_{12}^2} \begin{pmatrix} \frac{1}{m_{11}} & \frac{-m_{12}}{m_{11}m_{22}} \\ \frac{-m_{21}}{m_{11}m_{22}} & \frac{1}{m_{22}} \end{pmatrix} \tag{3.10}$$

where

$$r_{12}^2 = \frac{m_{12}^2}{m_{11}m_{22}} \quad .$$

Individual L.S. estimates are

$$\hat{\beta}_1 = \frac{1}{1-r_{12}^2}\left[\frac{m_{y1}}{m_{11}} - \frac{m_{12}m_{y2}}{m_{11}m_{22}}\right] = \frac{1}{1-r_{12}^2}\left[r_{y1}\sqrt{\frac{m_{yy}}{m_{11}}} - r_{12}r_{y2}\sqrt{\frac{m_{yy}}{m_{11}}}\right]$$

$$\hat{\beta}_2 = \frac{1}{1-r_{12}^2}\left[\frac{m_{y2}}{m_{22}} - \frac{m_{21}m_{y1}}{m_{11}m_{22}}\right] = \frac{1}{1-r_{12}^2}\left[r_{y2}\sqrt{\frac{m_{yy}}{m_{22}}} - r_{12}r_{y1}\sqrt{\frac{m_{yy}}{m_{22}}}\right] \quad .$$

If

$$\sqrt{\frac{m_{yy}}{m_{22}}} \doteq \sqrt{\frac{m_{yy}}{m_{11}}} \quad \text{and} \quad r_{12} \doteq 1, \text{ then } \hat{\beta}_1 \doteq -\hat{\beta}_2 \ .$$

Following Sastry (1970), $r_{y1} = r_{y2}$ when X_1 is proportional to X_2, it can be shown that

$$\lim_{r_{12}\to 1} \hat{\beta}_1 = \frac{r_{y1}}{2}\sqrt{\frac{m_{yy}}{m_{11}}}$$

and (3.11)

$$\lim_{r_{12}\to 1} \hat{\beta}_2 = \frac{r_{y2}}{2}\sqrt{\frac{m_{yy}}{m_{22}}}$$

Since $m_{11} = km_{22}$ when $r_{12} = 1$, where k is a scalar constant depending on the unit of measurement of X_1 and X_2 (without loss of generality, put $k = 1$), we find from (3.11) that the LS estimates of β_1 and β_2 tend to equal in magnitudes but opposite in sign as collinearity between X_1 and X_2 becomes perfect. Note that the regression sum of squares is

$$\begin{aligned} RSS = \hat{\beta}'C^{-1}\hat{\beta} &= \hat{\beta}_1^2 m_{11} + \hat{\beta}_2^2 m_{22} + 2\hat{\beta}_1\hat{\beta}_2 m_{12} \\ &= \left[\frac{r_{y1}^2}{4} + \frac{r_{y2}^2}{4} + \frac{r_{y1}r_{y2}}{2} r_{12} \right] m_{yy} \\ &\to r_{y1}^2 m_{yy} \quad \text{as} \quad r_{12} \to 1 \quad \text{and} \quad r_{y1} \to r_{y2} . \end{aligned}$$

Thus, the explanatory power of one regressor is to be divided up equally to two regressors if they are perfectly collinear.

On the other hand, the covariance matrix of the LS estimator is

$$\sigma^2 C = \frac{\sigma^2}{m_{11}m_{22}(1-r_{12}^2)} \begin{pmatrix} m_{22} & -m_{12} \\ -m_{12} & m_{11} \end{pmatrix} . \qquad (3.12)$$

As $r_{12}^2 \to 1$, then regardless of the sample size T (which affects m_{11} and m_{22}) and the error variance σ^2, the elements in the covariance matrix (3.12) will be large resulting in insignificant t-ratios unless the magnitudes of r_{yi} $(i = 1,2)$ are correspondingly large.[3] The power of the usual t-test is a direct function of the value of the non-centrality parameter

$$\lambda_i = \frac{m_{ii}\beta_i^2(1-r_{12}^2)}{\sigma^2} \qquad i = 1,2 \qquad (3.13)$$

---

[3] Johnston (1972) showed that the estimation of σ^2 is not seriously affected by collinearity so that large standard errors are adequate indicators of multicollinearity.

Given β_i, λ_i is clearly higher the higher the variation of X_i, the larger the sample size T, the smaller the error variance σ^2 and the lower the intercorrelation r_{12}^2. In a Monte Carlo study, Newhouse (1971) found that for $T \geq 25$, LS estimates yield a reasonably powerful test when $r_{12}^2 < .9$ and only when $T \leq 25$ and $r_{12}^2 > .99$ do LS estimates deteriorate such that they give wrong signs and low powers. However, his study was based on only two regressors with $\beta_1 = \beta_2 = 1$ and $R^2 = r_{y1}^2 + r_{y2}^2$ was set at above 90%. It is doubtful if the result would represent more complicated situations with higher signal-noise ratio.

For the general p (> 2) regressors case, it is more difficult to assess the effect of multicollinearity on individual LS estimates. It can be shown [Theil (1971)] that the diagonal elements of $(X'X)^{-1}$ are

$$C_{jj} = [m_{jj}(1-R_j^2)]^{-1} \qquad j = 1 \ldots p \tag{3.14}$$

where R_j^2 = coefficient of determination of regressing X_j on the other p-1 regressors.

Obviously, any pair-wise correlation coefficient r_{ij}^2 $(i \neq j)$ provides a lower bound for this quantity. Hence Klein's (1962) rule of thumb that $r_{ij}^2 > R^2$ to be a warning of multicollinearity is not informative in this case. In fact there are situations where $r_{ij}^2 = 0$ (such as seasonal dummies) but with perfect collinearity among regressors as a group.

The variance of individual LS estimates

$$\text{var}(\hat{\beta}_j) = \sigma^2[m_{jj}(1-R_j^2)]^{-1} \qquad j = 1 \ldots p \tag{3.15}$$

will be large when $R_j^2 \to 1$. The off-diagonal elements of $(X'X)^{-1}$ can be written as

$$C_{ij} = \frac{-m_{ij.(p-2)}}{\sqrt{m_{ii}(1-R_{i.(p-2)}^2)m_{jj}(1-R_{j.(p-2)}^2)}} \qquad i \neq j \tag{3.16}$$

where $m_{ij.(p-2)}$ = the partial cross product of X_i and X_j adjusted for the remaining $(p-2)$ regressors

$R^2_{i.(p-2)}$ = the coefficient of determination for the regression of X_i on the $(p-2)$ regressors excluding X_j.

Large magnitudes of the elements in the covariance matrix are implied by (3.15) and (3.16) when any of X_i and X_j is or both are involved with collinearity with other regressors.

The individual LS estimates are given by

$$\hat{\beta}_j = \sum_{i=1}^{p} C_{ij} m_{yi} \qquad j = 1 \ldots p \tag{3.17}$$

If X_j is involved in collinearity, these estimates will generally be large in absolute value irrespective of the sign of β_j $(j = 1 \ldots p)$ and the actual explanatory power of X_j. Thus a large estimated coefficient does not reflect the variable's predictive ability. The power of individual t-test is proportional to the noncentrality parameter

$$\lambda = \frac{\beta_j^2 (1-R_j^2) m_{jj}}{\sigma^2} \tag{3.18}$$

indicating that the power of hypothesis test based on LS estimates is impaired by multicollinearity (R_j^2 being close to 1).

It is often argued that [Theil (1961)] if the aim is to predict the dependent variable, the result will not be seriously affected by collinearity unless the correlation among the regressors in the sample period does not extend to the prediction period. This phenomenon is better analysed in a different co-ordinate system.

Let P be a $(T \times p)$ matrix such that $P'P = I_p$. Columns of P are principal co-ordinates of X.

G be a $(p\times p)$ orthogonal matrix. Columns of G are e-vectors of $X'X$.

$\Lambda = \text{diag}\{\lambda_1 \ldots \lambda_p\}$, where $\lambda_1 > \ldots > \lambda_p$ are e-roots of $X'X$.

Then, we can write

$$X = P\Lambda^{\frac{1}{2}}G' = \sum_{i=1}^{p} \sqrt{\lambda_i} p_i g_i' \tag{3.19}$$

and the LS estimates as

$$\hat{\beta} = (X'X)^{-1}X'y = G\Lambda^{-\frac{1}{2}}P'y = G\Lambda^{-\frac{1}{2}}r = G\hat{\gamma} \tag{3.20}$$

where r is $m_{yy}^{\frac{1}{2}}$ times the vector of simple correlations between the principal co-ordinates and the dependent variable.

$\hat{\gamma}$ is the vector of principal component estimates which is related to LS estimate by $G'\hat{\beta}$.

The smallest e-root λ_p identifies the e-vector that describes the direction where the data is the least informative. Multicollinearity in the data is thus reflected by the fact that Xg_p being close to zero. If the expected value of the dependent variables to be predicted is $\mathcal{E}(y|x_o) = x_o'\beta$, say , the LS predictor $x_o'\hat{\beta}$ has variance equal to

$$\begin{aligned} \text{var}(x_o'\hat{\beta}) &= x_o'(X'X)^{-1}x_o\sigma^2 \\ &= x_o' \left\{ \sum_{i=1}^{p} \frac{1}{\lambda_i} g_i g_i' \right\} x_o\sigma^2 . \end{aligned} \tag{3.21}$$

If collinearity structure remains the same in the prediction period, we have $x_o'g_p$ also close to zero, so that the term $(x_o'g_p)^2/\lambda_p$ in the sum at (3.21) will still be small despite the fact that λ_p is small. However, if collinearity structure changes then the LS predictor will be very unreliable because individual coefficients were not estimated

accurately. Malinvaud (1970, p.216) illustrated this phenomenon in a real data situation.

The fact that multicollinearity would do serious "harm" when we apply LS technique mechanistically has led econometricians to develop methods for its detection. Frisch's (1934) proposal of "bunch map" analysis which entails the computation of all possible regressions among the regressors is rarely practised because of its computational burden. A significant contribution in this direction was made by Farrar and Glauber (1967) who made use of information provided by the matrix $C = (X'X)^{-1}$. As noted earlier, the diagonal elements of C defined at (3.14) are related to the coefficients of determination of regressing one regressor on the rest. A better rule of thumb than Klein's suggestion is therefore that multicollinearity is deemed as "harmful" if $R_j^2 > R^2$. R^2 is the usual coefficient of determination of the LS regression. Marquardt (1970) called the diagonal elements C_{jj} $(j = 1...p)$ as "variance inflation factor" (VIF) and proposed that, for standardised data, a VIF > 5 implies that multicollinearity is responsible for poor LS estimates. Farrar and Glauber (1967) proposed various test statistics for the detection of the presence, localization and pattern of multicollinearity in a given data set. The tests are to be carried out in stages:-

(1) Test for the presence and severity of multicollinearity based on the approximate distribution of $\det(X'X)$ under the hypothesis of orthogonality.

(2) Test for the dependence of particular variables on other members of X based on the diagonal elements of C.

(3) Examine the pattern of interdependence among X based on the off-diagonal elements of C.

These tests, nevertheless, were criticised as being too "mechanistic" [see Leser (1968)] and suffering from the same problem as it proposed to solve [see Valentine (1969)]. The assumption that regressors are random drawings from multivariate independent normal distributions is also unrealistic. The often used method of checking inconsistencies in results of hypothesis testing to detect multicollinearity is also ambiguous as such conflict could also arise with orthogonal regressors [see Geary and Leser (1968)]. Furthermore, a model can always be reparameterised so that the regressors are orthogonal, although such transformation may lack economic justification. Experience also shows that $\det(X'X)$ being small need not create serious estimation problem as it depends on the unit of measurement of the data. A more reasonable measure is the conditioning number represented by the ratio of the largest to the smallest e-root.

Accepting that collinearity is a fact of life, the more pressing concern for the econometricians who handle non-experimental time series data is to find a solution to such sticky problem rather than to detect its existence. Traditional solutions are largely of an ad hoc nature. The dropping of regressor variables or principal components[4] [Massy (1965)] are simply ad hoc ways of incorporating exact a priori restrictions on the parameter space. This is susceptible to model misspecification, consequences of which have been discussed in §2.2. A slightly more flexible alternative is to allow for some random variation on the restriction and to incorporate it in a manner proposed by Theil and Goldberger (1961). These procedures are tantamount to the informal use of prior information, which, in fact, could be handled more conveniently and systematically by

---

[4] Extensions to "Generalised Inverse" method which allows the deletion of a fraction of the principal components and "Latent Root" method which analyses the e-roots of $[y\ X]'[y\ X]$ rather than simply $X'X$ have been considered in the literature. [See Marquardt (1970), Webster et al (1974) respectively].

Bayesian techniques. However, one later development of a non-Bayesian solution to the collinearity problem was proposed by Hoerl and Kennard (1970). Their ridge regression, presumably aimed at solving collinearity problem as well as improving the mean square error property of the LS estimator, has aroused considerable controversy in the profession. Despite its lack of theoretical underpinning, the ridge estimator appears to work tremendously well in certain situations [see Dempster et al (1977) and Lawless (1978)]. We shall discuss the theoretical and practical aspects of the ridge regression technique in the next two sections. In section 3.6, we present a unified treatment of all the biased estimators in a general class of estimators, all of which are sort of minimum mean-square-error estimators (MMSE) under certain restrictions. In the last section, a Bayesian interpretation of the multicollinearity problem is furnished.

§3.3 The "Ridge" Existence Theorems[5]

The innocent-looking ordinary ridge estimator (ORE) of Hoerl and Kennard (1970) has been received by the profession with mixed feelings. Some despise the technique as nothing more than a simple trick of numerical analysis lacking statistical or economic justification [e.g. Coniffe and Stone (1973)]. Others find the technique helpful in obtaining "sensible" estimates under suitable conditions when all other classical techniques fail [Vinod (1974), Brown and Payne (1975)]. The ridge estimator conditional on a given ridge parameter k (k being a positive scalar) is

$$\begin{aligned}\hat{\beta}_{(k)} &= (X'X + kI_p)^{-1}X'y \\ &= Z\hat{\beta}\end{aligned} \tag{3.22}$$

where

---

[5] Part of this section is reported in Lee (1977) and Lee (1978a).

$$Z = (X'X + kI_p)^{-1}X'X$$

$\hat{\beta}$ is the unrestricted LS estimator

$$= (X'X)^{-1}X'y \ .$$

Hoerl and Kennard motivated the ordinary ridge estimator (ORE) by pointing out that the expected length of the usual LS estimator

$$\mathcal{E}(\hat{\beta}'\hat{\beta}) = \beta'\beta + \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i} \tag{3.23}$$

is bounded below by $\beta'\beta + \frac{\sigma^2}{\lambda_p}$. λ_p being the smallest e-root of $X'X$ will be small when the data are collinear giving coefficient estimates whose absolute values are too large, possibly with "wrong" sign. By restricting the length of the estimator within a sphere of finite radius, centered at the origin, the resulting LS estimator subject to this constraint can be shown to be the ORE at (3.22) with k interpreted as the Lagrange multiplier of the optimal solution. The mean and variance of the ORE, at a fixed k, are respectively

$$\mathcal{E}\hat{\beta}_{(k)} = Z\beta \tag{3.24}$$

and

$$\text{var } \hat{\beta}_{(k)} = \sigma^2 Z(X'X)^{-1}Z' \ . \tag{3.25}$$

Obviously, $\hat{\beta}_{(k)}$ is a biased estimator of β, the bias being $(Z-I)\beta$ and the risk matrix measure of $\hat{\beta}_k$ is,

$$MSE(\hat{\beta}_{(k)}) = \sigma^2 Z(X'X)^{-1}Z' + (Z-I)\beta\beta'(Z-I)' \tag{3.26}$$

Theorem 3.3.1 (Hoerl and Kennard (1970))

There always exists a $k > 0$ such that the mean square error risk of $\hat{\beta}_{(k)}$ is strictly less than that of $\hat{\beta}$. The upper bound for such

a k is $\frac{\sigma^2}{\gamma_{max}^2}$ denoted by k^*_{HK}.

Proof. Total mean square error risk of $\hat{\beta}_k$ is given by

$$R_I(\hat{\beta}_k) = \text{tr MSE}(\hat{\beta}_k) = \sigma^2 \text{tr } Z(X'X)^{-1}Z' + k^2\beta'(X'X+kI)^{-2}\beta \qquad (3.27)$$

$$= \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+k)^2} + k^2 \sum_{i=1}^{p} \frac{\gamma_i^2}{(\lambda_i+k)^2}$$

$$= \phi_1(k) + \phi_2(k)$$

where $\phi_1(\cdot)$ and $\phi_2(\cdot)$ are the sum of variances and sum of squared bias respectively. Since $\hat{\beta}_{(k)} = \hat{\beta}$ at $k = 0$, it is sufficient to prove that there always exists a $k^* > 0$ such that $R_I(\hat{\beta}_{(k)})$ is a decreasing function of k in the region $[0,k^*]$

$$\frac{dR_I(\hat{\beta}_{(k)})}{dk} = -2\sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+k)^3} + 2k \sum_{i=1}^{p} \frac{\lambda_i\gamma_i^2}{(\lambda_i+k)^3} \qquad (3.28)_1$$

$$\leq -2\sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+k)^3} + 2k\gamma_{max}^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+k)^3} \qquad (3.28)_2$$

where γ_{max}^2 = the largest squared element of the true Principal Component vector $\gamma = G'\beta$.

It is clear that the right side of $(3.28)_2$ is negative if and only if $\sigma^2 > k\gamma_{max}^2$. Thus the region of k where ORE has total mean square error decreasing in k is $[0, \frac{\sigma^2}{\gamma_{max}^2}]$. This region is nonempty so far as the elements of γ are bounded from above. □

Remark 1. This theorem is of little practical help to econometricians as the improvement region depends on the unknown parameters. Of course a value of k close enough to zero is very likely to improve upon OLS. Nevertheless, such a small k would give an almost identical estimate

as the OLS which will hardly have any significant improvement to justify a biased estimator. If the unknown is replaced by some estimates (say the OLS or iterative ridge estimates) then the proof in the above theorem is invalid.

Remark 2. Critics may argue against the use of the squared error loss which gives equal weight to different components, and comparison is only made in aggregate. It is possible that some components are estimated with bigger risk than the corresponding LS components. However, this possibility is guarded against by the following result.

Theorem 3.3.2 (Theobald (1974))

There exists $k > 0$ such that the difference of the risk matrix measures of the OLS and that of the ORE is positive semidefinite. The upper bound for such k is $\frac{2\sigma^2}{\beta'\beta}$ denoted by k^*_{THE}.

Proof.

$$\begin{aligned} MSE(\hat{\beta})-MSE(\hat{\beta}_{(k)}) &= \sigma^2(X'X)^{-1}-\sigma^2Z(X'X)^{-1}Z'-(Z-I)\beta\beta'(Z-I)' \\ &= \sigma^2(X'X+kI)^{-1}\{(X'X+kI)(X'X)^{-1}(X'X+kI)-X'X\}(X'X+kI)^{-1} \\ &\qquad -k^2(X'X+kI)^{-1}\beta\beta'(X'X+kI)^{-1} \\ &= k(X'X+kI)^{-1}\{\sigma^2[2I+k(X'X)^{-1}]-k\beta\beta'\}(X'X+kI)^{-1} . \end{aligned} \tag{3.29}$$

The right side is positive semidefinite if and only if

$$\sigma^2[2I+k(X'X)^{-1}] - k\beta\beta' \tag{3.30}$$

is positive definite. This is satisfied if

$$\text{(i)} \quad k < \frac{2\sigma^2}{\beta'\beta} \tag{3.31}$$

or

$$\text{(ii)} \quad \beta'X'X\beta < \sigma^2 . \tag{3.32}$$

Remark 1. The second condition is independent of k, which in fact is comparable to the Toro-Wallace strong criterion for the null vector to dominate OLS.

Remark 2. The improvement region of Theobald is seen to be much narrower than that of Hoerl and Kennard which is at most $\frac{p}{2}$ times the former. It is clear that if the elements of β are unbounded, both regions degenerate at the origin. This conforms with Barnard's (1963) result that the minimum mean square error criterion in this circumstance leads immediately to the least square procedure.

COROLLARY 3.3.2 (Farebrother (1976))

There exists $k > 0$ such that the difference of the risk matrix measures of OLS and that of ORE is positive semidefinite. A necessary and sufficient condition is that

$$\beta'[\frac{2}{k} I_p + (X'X)^{-1}]^{-1}\beta < \sigma^2 . \tag{3.33}$$

Proof. Obvious from (3.30) and the fact that $k > 0$. ■

COROLLARY 3.3.3 (Swindle (1976))

For the difference of the risk matrix measure of OLS and that of ORE to be positive semidefinite, it is necessary and sufficient that k is in the open interval

$$I(\beta,\sigma^2,X) = \begin{cases} (0, \infty) & \text{if } \beta'X'X\beta < \sigma^2 \\ (0, -\frac{2}{\psi}) & \text{otherwise} \end{cases} \tag{3.34}$$

where ψ = minimum e-root of $(X'X)^{-1} - \frac{\beta\beta'}{\sigma^2}$ which is negative definite

$= d_s - \frac{\beta'\beta}{\sigma^2}$, $d_s = \frac{1}{\lambda_1}$ is the smallest e-root of $(X'X)^{-1}$.

Proof. For (3.30) to be positive definite, the case when $\beta'X'X\beta < \sigma^2$ is trivially established in (3.32).

Suppose that $(X'X)^{-1} - \frac{\beta\beta'}{\sigma^2}$ is negative definite. Then (3.30) is bounded below by

$$2I_p + \psi k I_p \ . \tag{3.35}$$

If this lower bound is positive definite, then (3.30) is also positive definite. Hence, when $k < -\frac{2}{\psi}$, ORE would dominate OLS if $\beta'X'X\beta > \sigma^2$. □

Note: in the proofs of ridge existence theorem above, the vector β can be replaced everywhere by $\beta - \beta_o$, if the ORE at (3.22) is changed to be

$$\hat{\delta}_{(k)} = \hat{\beta}_{(k)} - \beta_o = Z(\hat{\beta}-\beta_o) \ . \tag{3.36}$$

Returning to the mean square error risk $R_I(\cdot)$ we can obtain, by setting $(3.28)_1$ to zero, an optimal value of the ridge parameter. It turns out to be a fixed point solution of

$$\sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+k)^3} = k \sum_{i=1}^{p} \frac{\lambda_i\gamma_i^2}{(\lambda_i+k)^3} \ . \tag{3.37}$$

There is no explicit solution to (3.37) except in the special case

(a) $\gamma_1^2 = \ldots = \gamma_p^2 = \bar{\gamma}^2$ so that $k^* = \frac{\sigma^2}{\bar{\gamma}^2}$

or

(b) all $\lambda_i = 0$ except one

or

(c) all $\lambda_i = 1$, $i = 1 \ldots p$, (an orthogonal design on standardised data) then the k minimising (3.37) is $\frac{p\sigma^2}{\beta'\beta}$ or $\frac{p\sigma^2}{\gamma'\gamma}$.

If we allow k to assume different values corresponding to different e-roots, then the generalised ridge estimator is

$$\hat{\beta}_{GRE} = (X'X + GKG')^{-1}X'y \tag{3.38}$$

where $K = \text{diag}\{k_1 \ldots k_p\}$

G has as its columns the e-vectors of $X'X$.

$R_I(\hat{\beta}_{GRE})$ is minimised when $k_i = \frac{\sigma^2}{\gamma_i^2}$ $(i = 1 \ldots p)$ as given by Hoerl and Kennard (1970) and Goldstein and Smith (1974).

Note that the harmonic mean of these "optimal" k_i is

$$\frac{p}{\frac{1}{k_1} + \frac{1}{k_2} + \ldots + \frac{1}{k_p}} = \frac{p}{\sum_{i=1}^{p} \frac{\gamma_i^2}{\sigma^2}} = \frac{p\sigma^2}{\gamma'\gamma} = \frac{p\sigma^2}{\beta'\beta} \tag{3.39}$$

which is seen to be independent of the e-roots of $X'X$.

The existence theorems are essentially a feature of the characteristics of the two functions $\phi_1(\cdot)$ and $\phi_2(\cdot)$ which form the mean squared error risk $R_I(\hat{\beta}_{(k)})$ in (3.27). It is easy to check that the sum of variances, $\phi_1(k)$, is a monotonically decreasing function of k while the sum of squared bias, $\phi_2(k)$, is a monotonically increasing function of k. The fact that the gradient of $\phi_1(k)$ has a magnitude greater than that of $\phi_2(k)$ at the origin implies that there must be a positive value of k close enough to the origin such that the introduction of a small bias is more than compensated by a reduction in variance. Unfortunately, the bias term is necessarily dependent on the location of the unknown true value of the coefficient vector resulting in ambiguity of delimiting the upper limit of the "improvement" region for k. Interestingly, this result of the ridge estimator does not depend in any way on the collinearity structure of data. Leamer (1977) emphasized these points humorously by suggesting that there always exists a "valley"

estimator obtained through adding a small positive scalar k to the off-diagonal elements of the $X'X$ matrix, provided $\hat{\beta}$ lies in the first orthant, which would dominate OLS.

In order to have a better idea of how the location of true coefficient vector would affect the risk of the ordinary ridge estimator, it suffices to consider only the sum of squared bias function, $\phi_2(k)$.

Theorem 3.3.3

For $k > 0$, $\lambda_1 > \ldots > \lambda_p$ e-roots corresponding to the e-vectors $\underset{\sim}{g}_1 \cdots \underset{\sim}{g}_p$ the sum of squared bias function $\phi_2(k)$ are bounded in the following region

$$\left(\frac{k}{\lambda_1+k} \right)^2 \beta'\beta \leq \phi_2(k) \leq \left(\frac{k}{\lambda_p+k} \right)^2 \beta'\beta \ . \tag{3.40}$$

The first equality holds if β lies in the direction of $\underset{\sim}{g}_1$ (the e-vector corresponding to the largest e-root λ_1) whereas the second equality holds if β lies in the direction of $\underset{\sim}{g}_p$ (the e-vector corresponding to the smallest e-root λ_p).

Proof. Let us write $\beta'\beta = \sum_{i=1}^{p} \gamma_i^2$, where γ_i $(i = 1 \ldots p)$ are the true principal components of β.

For any $k > 0$

$$\left(\frac{k}{\lambda_1+k} \right)^2 \gamma_j^2 \leq \left(\frac{k}{\lambda_j+k} \right)^2 \gamma_j^2 \qquad j = 1 \ldots p \ .$$

Adding up these p inequalities, we have

$$\left(\frac{k}{\lambda_1+k} \right)^2 \sum_{j=1}^{p} \gamma_j^2 \leq \sum_{j=1}^{p} \left(\frac{k}{\lambda_j+k} \right)^2 \gamma_j^2 = \phi_2(k) \ . \tag{3.41}$$

If β lies in the direction of $\underset{\sim}{g}_1$, then β is orthogonal to $\underset{\sim}{g}_2, \ldots, \underset{\sim}{g}_p$ such that $\beta'\beta = \gamma'\gamma = \gamma_1^2$ $(\because \gamma_i = \beta'\underset{\sim}{g}_i \quad i = 1 \ldots p)$ and

$\gamma_i^2 = 0 \quad (i \neq 1)$. The equality in (3.41) holds in this situation and the ridge estimator will attain the minimum bias. Of course, the bias will be zero if $k = 0$ but then $k = 0$ is not the optimal one in terms of risk, due to the existence theorems.

Similarly, we use the fact that

$$\left(\frac{k}{\lambda_p + k} \right)^2 \gamma_j^2 \geq \left(\frac{k}{\lambda_j + k} \right)^2 \gamma_j^2 \qquad j = 1 \ldots p$$

to yield

$$\left(\frac{k}{\lambda_p + k} \right)^2 \sum_{j=1}^{p} \gamma_j^2 \geq \sum_{j=1}^{p} \left(\frac{k}{\lambda_j + k} \right)^2 \gamma_j^2 = \phi_2(k) \ . \tag{3.42}$$

If β lies in the direction of $g_{\sim p}$, then β is orthogonal to $g_{\sim 1} \cdots g_{\sim p-1}$ such that $\beta'\beta = \sum_{j=1}^{p} \gamma_j^2 = \gamma_p^2$ and $\gamma_i^2 = 0 \quad (i \neq p)$. Thus equality in (3.42) holds in this situation and the ridge estimator will attain the maximum bias.

Combining (3.41) and (3.42) we have established Theorem 3.3.3. □

As noted by Baldwin and Hoerl (1978), the implication of this result is that the potential of risk reduction of the ORE from that of OLS depends on the bearing of β in the system of principal co-ordinates and its length. If $\beta'\beta$ is *less* than $R_I(\hat{\beta}) = \text{var}(\hat{\beta}) = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j}$, then any value of k will give ridge estimates superior to least squares estimates, because $R_I(\hat{\beta}_{(k)})$ approaches $\beta'\beta$ as k goes to infinity. However, if $\beta'\beta$ is substantially larger than $R_I(\hat{\beta})$, then there is some value of k that would cause the ridge estimate to be excessively biased so that the resulting risk of ORE could be substantially higher than that of OLS.

In economics, most of the time-series data are positively correlated so that $X'X$ has positive off-diagonal elements. Imagine the approximate

situation where X is the $(T\times p)$ data matrix of a distributed lag model of a single explanatory variable which is positively autocorrelated, then $X'X$ is the $(p\times p)$ matrix of sample autocorrelation functions. Anderson (1971) showed that asymptotically the e-vectors for such matrix are alternately $\frac{1}{2}(\underset{\sim}{u}_s + \underset{\sim}{u}_{p-s})$ and $\frac{1}{2i}(\underset{\sim}{u}_s - \underset{\sim}{u}_{p-s})$, where $i = \sqrt{-1}$ and $\underset{\sim}{u}_s$ is the $(p\times 1)$ vector whose t'th element is $e^{i\frac{2\pi ts}{p}}$ $(s = 0,\ldots,\frac{p}{2})$. Since positively autocorrelated series has most of its spectral power concentrated around the low frequency points (i.e. small s), it is not hard to see that the e-vector corresponding to the largest e-root is when $s = 0$, i.e.

$$\underset{\sim}{g}_1 = \tfrac{1}{2}(u_o + u_p) = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \tag{3.43}$$

This may explain why ridge estimator has been applied with success in distributed lag models when the true elements of β are small positive values and all of which are approximately equal [see Trivedi and Lee (1979)]. We shall see in the last section of this chapter that a Bayesian interpretation of ridge estimator is justified when the assumption of "exchangeability" holds. Turn the argument around, if we want to improve on the LS estimate by a ridge estimate, we might as well first examine the e-vector of $X'X$ corresponding to the largest e-root to give us some hint as to the direction in which the true coefficient vector should lie in order that our attempt will be fruitful.

§3.4 Selection of the Ridge Parameter and the Recursive Algorithm

As there is a one to one relationship between the ridge parameter k and the ridge estimate of regression coefficients $\hat{\beta}_{(k)}$, the problem of choosing the best ridge estimate is the same as deciding which ridge

parameter to be selected. Since the analytically "optimal" ridge parameter is a function of the unknown coefficients and error variance, this remains a difficult problem. Furthermore, if k is chosen as a function of the dependent variable y, then the sampling properties of the resulting ridge estimator remains unknown. Nevertheless, Monte Carlo studies and actual data analyses indicate that some rules of selecting k appear to live up to the desirable properties promised by a known "optimal" k. Since the ridge estimator is discovered, there have been numerous different rules proposed to select the ridge parameter. Some based on statistical argument, many based on heuristic rules of thumb. We shall discuss some of the more popular ones under two broad categories[6], namely control oriented and prediction oriented selection criteria.

(A) Control Oriented Criteria

(1) Hoerl and Kennard (1970)'s ridge trace technique is possibly the best known. This is to select k graphically at a point where all the elements of the vector of ridge estimate begin to stabilise or some wrongly signed elements have a correct sign. This technique is informative but lacking uniqueness. Different users may identify different points of stable estimates along the trace, especially since the ridge parameter is dependent on units of measurement of the variables. Although Hoerl and Kennard advised users to standardise their data, taking such a step effectively alters the unweighted loss function to a weighted one which may not be desirable. Finally, the selected k is not a data-dependent variable so that the sampling properties of ridge estimator are affected.

---

[6] These two categories are considered because they are the commonly used criteria in econometrics to gauge the goodness of an estimator. "Control" is referred to the estimation of individual coefficients (partial effects) whereas "Prediction" is referred to the estimation of the endogenous variable.

(2) Obenchain (1975) proposed calculating the sum of squares of all $\binom{p}{2}$ correlations between the coefficients of the ridge estimates, abbreviated SSCBC. This measure depends on the X matrix only and hence is not stochastic so that the sampling results of ridge estimate may be preserved. However, SSCBC criterion very often leads to a choice of k which is substantially larger than the "stable" choice indicated by the trace and produces a biased estimate lying exterior to a conventional confidence region. If one tries to control the amount of shrinkage within certain confidence region, then again the resulting estimator's sampling property is affected.

(3) Vinod's (1976) Index of Stability of Relative Magnitudes (ISRM) defined by

$$\text{ISRM} = \sum_{i=1}^{p} \left[\frac{p\delta_i}{\sum_{j=1}^{p} \frac{\delta_j^2}{\lambda_j}} - 1 \right]^2$$

where

$$\delta_i = \frac{\lambda_i}{\lambda_i + k} \qquad i = 1 \ldots p$$

which has a minimum value of zero when the design is orthogonal [because then $\lambda_i = 1$ $(i = 1 \ldots p)$. $\delta_i = \frac{1}{1+k} = \bar{\delta}$, all i, so that $p\bar{\delta} = \sum_{j=1}^{p} \delta_j = \frac{p}{1+k}$]. This criterion is also subject to the same criticisms as SSCBC. Simulation results of Wichern and Churchill (1978) indicated this rule leads to an inferior ORE than other rules.

(4) Marquardt's (1970) rule of thumb of controlling the amount of bias by restricting the VIF's, diagonal elements of the covariance matrix of the ridge estimator, to be between [1, 10] is also too subjective.

(5) McDonald-Galarneau (1973) suggested control the length of the estimated vector so that k is a solution from equating $|\hat{\beta}_{(k)}|^2$ to a pre-assigned value of "correct length" of β. The "correct length" is estimated by an unbiased estimate based on least squares as

$$Q = \hat{\beta}'\hat{\beta} - \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i} \quad . \tag{3.45}$$

Wichern and Churchill (1978) found that this criterion leads to a better performance of ridge estimates in terms of mean squared errors than those selected by other criteria in their study.

(B) Prediction Oriented Criteria

(1) The best known and widely referred to criterion under this category is perhaps Mallow's $C(\hat{\beta}_{(k)})$ which is an estimator of the scaled mean prediction error, $R_{XX}(\hat{\beta}_{(k)})$.

$$C(\hat{\beta}_{(k)}) = (T-p-1) \frac{ESS(\hat{\beta}_{(k)})}{ESS(\hat{\beta})} - T + 2 \sum_{i=1}^{p} \left(\frac{\lambda_i}{\lambda_i+k} \right) + 2 \tag{3.46}$$

where $ESS(\hat{\beta})$ = error sum of squares of $\hat{\beta}$. Experience shows that this criterion often leads to a choice of k very near the origin. Plotting $C(\hat{\beta}_{(k)})$ and $ESS(\hat{\beta}_{(k)})$ reveals their similarity. As is well known $ESS(\hat{\beta}_{(k)})$ is minimised at $k = 0$.

(2) The often mentioned but seldom used technique is cross-validation which exhaustively re-estimates the coefficients when one observation is omitted and predicted at a time. The average of all such prediction squared errors of the omitted observation gives Allen's (1974) $PRESS(\hat{\beta}_{(k)})$ statistic. k is chosen such that PRESS is minimised. This technique may prove to be too expensive and time consuming even with today's computer technology, especially if the number of observations is

moderately large. Wahba (1976) suggested the technique to be applied to an orthogonal transformation of data with more than one observation omitted at a time. Of course this suggestion requires even greater computation and it is doubtful whether the effort could be justified.

(3) A selection criterion based on a likelihood approach was suggested by Obenchain (1975). The idea was that the likelihood function $L(\hat{\beta}_{(k)})$ of k_i based on an unrestricted p-parameter $(k_1 \ldots k_p)$ family should be much higher than that based on a restricted family of one or two parameters, such as $k_i = k\lambda_i^q$ which depends on only 2 parameters (k,q). Conditional on q, k can be chosen such that the statistic $-2\ell nL(\hat{\beta}_{(k)})$ is minimised. Note that $q = 0$ gives the Hoerl and Kennard's (1970) ordinary ridge family, $q = 1$ the Mayer and Willke's (1973) shrunken estimator and $q = 1-m$, (m an integer), the Goldstein and Smith (1974) family. Hoerl and Kennard (1975) pointed out that the optimal generalised ridge parameters $k_i = \frac{\sigma^2}{\gamma_i^2}$ which are inversely proportional to the regression coefficients and not to the e-roots, it appears that exploration of values of q other than zero will not be fruitful. Hence, Obenchain's $-2\ell nL(\hat{\beta}_{(k)})$ criterion is of limited value.

(4) Other criteria based directly on the theoretically optimal ridge parameter are also suggested. Hoerl and Kennard (1970) estimated k^* by $\hat{\sigma}^2/\hat{\gamma}_{max}$ using least squares estimates. Later, Hoerl, Kennard and Baldwin (1975) modified the estimate to be $p\hat{\sigma}^2 \Big/ \sum_{i=1}^{p} \hat{\gamma}_i^2$ and found that the latter performed well in their simulation study. However, Lawless and Wang (1976) reinterpreted their selection from a Bayesian viewpoint and recommended yet another modification of the estimate to be $p\hat{\sigma}^2 \Big/ \sum_{i=1}^{p} \lambda_i \hat{\gamma}_i^2$. Similar simulations done by Lawless and Wang (1976) favoured their modification. Further confusion about the relative merits

of the two estimates were added by the result of Wichern and Churchill (1978) who reported the poor performance of both rules.

(5) One final rule worth mentioning is the one proposed by Swamy, Mehta and Rappoport (1978). They proved that the LS estimation of σ^2 and $X\beta$ is quite accurate even under multicollinearity. As the strong condition for $\hat{\beta}_{(k)}$ to dominate $\hat{\beta}$ is given in (3.33), the dividing line

$$\beta'[\frac{2}{k} I_p + (X'X)^{-1}]^{-1}\beta = \sigma^2 \tag{3.47}$$

can be written as [see Rao (1973), p.33]

$$\beta'X'[\frac{2}{k} XX' + I]^{-1}X\beta = \sigma^2 \ . \tag{3.48}$$

Then, k_{SMR} can be obtained as the implicit solution of

$$\hat{\beta}'X'[\frac{2}{k} XX' + I]^{-1}X\hat{\beta} = \hat{\sigma}^2 \ . \tag{3.49}$$

The solution is unique as the left side is monotonic increasing in k.

In a simulation study, they found that mean square error of individual estimates based on this choice of k are uniformly smaller than that of the McDonald-Galarneau's (1975) "correct length" choice, which in turn is smaller than OLS. But this estimate performs less as well as Hoerl, Kennard and Baldwin's choice of $k = \frac{p\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}}$, which in turn was beaten by one using $k = p \times k_{SMR}$.

In summary, there is as yet no consensus as to which selection rule should be recommended to the user of ridge estimator. Some are good for prediction purposes while others are good for control purposes. Nevertheless, some rules emerge to be quite robust to the data structure and quite economical to apply without involving subjective decision. Since the sampling properties of ridge estimator will be seriously affected, the only evidence to favour a particular rule is purely based on Monte Carlo

studies whose limitations are universally recognised.

One special problem in connection with the selection process is the amount of computation required. All rules require calculation of various statistics corresponding to the ridge estimates at different values of k (including those suggested in B(4) and B(5) above if an iterative search for k^* as suggested by Lindley and Smith (1972) is carried out). Plackett's (1950) recursive algorithm can be shown to be of practical advantage. The updating formula for OLS on receipt of a new row of observations at time (T+1), say, [y(T+1), x(T+1)'], is

$$\hat{\underset{\sim}{\beta}}_{T+1} = \hat{\underset{\sim}{\beta}}_{T} + \frac{(X'X)^{-1}x(T+1)(y(T+1)-x(T+1)'\hat{\underset{\sim}{\beta}}_{T})}{1+x(T+1)'(X'X)^{-1}x(T+1)} \tag{3.50}$$

and the error sum of squares can be updated as

$$ESS_{T+1} = ESS_{T} + \frac{(y(T+1)-x(T+1)'\hat{\beta}_{T})^{2}}{1+x(T+1)'(X'X)^{-1}x(T+1)} \tag{3.51}$$

where $\hat{\beta}_T$ is LS estimate based on the original T observations

y(T+1) is the newly added element to the original T×1 vector y

x(T+1)' is the newly added row to the original T×p matrix X.

If there is an additional n observations represented by $[\underset{\sim}{w}\ Z]$ where $\underset{\sim}{w}$ is n×1 vector of new observations on the dependent variable and Z is n×p matrix of new observation on the regressors, then the updating formula for the LS estimates is

$$\hat{\beta}_{T+n} = \hat{\beta}_{T} + (X'X)^{-1}Z'(I_n+Z(X'X)^{-1}Z')^{-1}(\underset{\sim}{w}-Z\hat{\beta}_{T}) \tag{3.52}$$

and the error sum of squares can be updated as

$$ESS_{T+n} = ESS_{T} + (w-Z\hat{\beta}_{T})'[I+Z(X'X)^{-1}Z']^{-1}(w-Z\hat{\beta}_{T}) \ . \tag{3.53}$$

Applying these formulae to the ridge trace for various values of k, we can imagine the new data are in the form of p observations $[0, \sqrt{k}I_p]$

$$\hat{\beta}_{(k)} = \hat{\beta} - k(X'X)^{-1}(I_p + k(X'X)^{-1})^{-1}\hat{\beta} \ . \tag{3.54}$$

This formulation makes use of the originally computed $(X'X)^{-1}$ so that the need for storage of $X'X$ can be dispensed with.

§3.5 Relation of Ridge Estimators to Other Biased Estimators

The development of this section is based on a technique of Hocking et al (1976) who considered a class of biased estimators that includes all the estimators so far discussed in this thesis as special cases. These are basically minimum mean square error estimators subject to various forms of restrictions. In this framework, it is hoped that econometricians may be able to see the difference among various estimators so that an educated choice can be made on them according to the problem's requirements.

It proves to be more insightful to consider a canonical transform of the model and the various estimators in terms of the principal component estimates $\hat{\gamma}$ which are related to the least squares estimates $\hat{\beta}$ as

$$\hat{\beta} = G'\hat{\gamma} \ . \tag{3.55}$$

Define the new basis vectors

$$\hat{\theta}^{(i)} = \begin{pmatrix} \hat{\gamma}_1 \\ \hat{\gamma}_2 \\ \vdots \\ \hat{\gamma}_i \\ 0 \\ \vdots \\ 0 \end{pmatrix} \qquad i = 1 \ldots p \ . \tag{3.56}$$

a form of the truncated principal components estimates.

Then it is easy to see that the conventional principal component estimator (PCE) with an integer assigned rank $r(< p)$ is simply

$$\tilde{\xi}_{PCE} = \hat{\theta}^{(r)} \qquad r < p \tag{3.57}$$

while the full principal component estimator (the least squares estimate in the canonical form) is

$$\hat{\xi}_{LS} = \hat{\theta}^{(p)} = \hat{\gamma} \ . \tag{3.58}$$

Marquardt's generalised inverse estimator (GIE) or the "fractional" rank estimator is motivated by the fact that the $p-r$ components being truncated in (3.57) may contain some information about the true coefficients. In order to recapture this information, Marquardt (1970) suggested that if a fraction of the $(r+1)$'th component is retained as well, then the resulting estimator would solve the singularity problem without losing too much information. That is

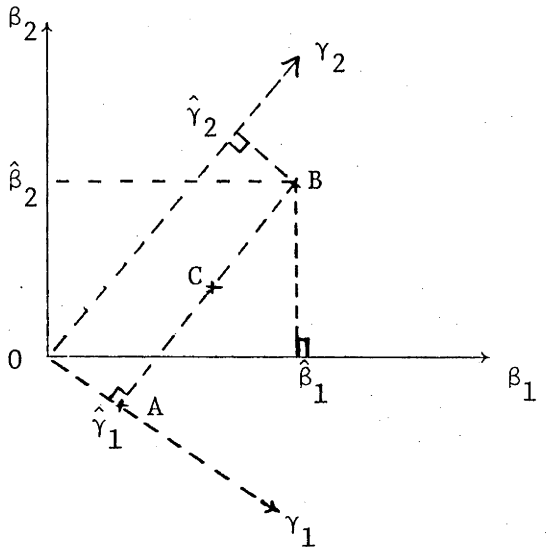
$$\tilde{\xi}_{GIE} = \hat{\theta}^{(r)} + a(\hat{\theta}^{(r+1)} - \hat{\theta}^{(r)}) \tag{3.59}$$

where $0 < a < 1$.

Note that $\tilde{\xi}_{GIE}$ differs from $\tilde{\xi}_{PCE}$ by the $(r+1)$'st element (which is $a\hat{\gamma}_{r+1}$) only.

If $p = 2$ and $r = 1$, $a = \frac{1}{2}$, then graphically, the estimators (3.57), (3.58), (3.59) can be represented by points A, B and C respectively.

FIGURE 3.1 : Principal Components



A = point of PCE
B = point of OLS
C = point of GIE

Let the general estimator be

$$\tilde{\xi} = \sum_{i=1}^{p} a_i \hat{\theta}^{(i)} \tag{3.60}$$

where a_i $(i = 1 \ldots p)$ are scalars to be determined. That is, the "general" estimator is a weighted average of all the new 'basis' vectors (or points of restricted principal components estimators). In view of the definition of $\hat{\theta}^{(i)}$, (3.60) can also be written as

$$\tilde{\xi} = \begin{pmatrix} \sum_{i=1}^{p} a_i & & & \\ & \sum_{i=2}^{p} a_i & & \\ & & \ddots & \\ & & & a_p \end{pmatrix} \begin{pmatrix} \hat{\gamma}_1 \\ \\ \\ \hat{\gamma}_p \end{pmatrix} = D\hat{\theta}^{(p)} = D\hat{\gamma} \tag{3.61}$$

where D is the diagonal matrix with j'th diagonal element given by $d_j = \sum_{i=j}^{p} a_i$ $(j = 1 \ldots p)$. Since $\hat{\theta}^{(i)}$ $(i = 1 \ldots p)$ are linearly independent, $\tilde{\xi}$ in (3.61) lies anywhere in the p-dimensional space.

Now, under square error loss, the risk of $\tilde{\xi}$

$$R_I(\tilde{\xi}) = \mathcal{E}(\tilde{\xi}-\xi)'(\tilde{\xi}-\xi)$$

$$= tr\sigma^2 D\Lambda^{-1}D + tr(D-I)\gamma\gamma'(D-I)$$

$$= \sigma^2 \sum_{i=1}^{p} \frac{d_i^2}{\lambda_i} + \sum_{i=1}^{p} \gamma_i^2 (d_i-1)^2 \ . \qquad (3.62)$$

(1) If we minimise (3.62) with respect to d_i's subject to the constraints that

$$d_i = \text{constant} \qquad i = 1,\ldots,p$$

then, the optimal weight is

$$d_i^* = \frac{\gamma'\gamma}{\gamma'\gamma+\sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i}} \qquad i = 1,\ldots,p \ . \qquad (3.63)$$

This leads to the shrunken estimator of Mayer and Willke (1973)

$$\tilde{\xi}_{SH} = \frac{\gamma'\gamma}{\gamma'\gamma+\sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i}} \hat{\gamma} = \left(1 - \frac{\sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i}}{\gamma'\gamma+\sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i}} \right) \hat{\gamma} \qquad (3.64)$$

If we estimate the unknown $\dfrac{\sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i}}{\gamma'\gamma+\sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i}}$ by $\dfrac{p-2}{\hat{\gamma}'\hat{\gamma}}$, then we have the Stein-James estimator

$$\tilde{\xi}_{SJ} = \left(1 - \frac{p-2}{\hat{\gamma}'\hat{\gamma}} \right) \hat{\gamma} \ .$$

(2) The Generalised Inverse estimator (3.59) is obtained by minimising (3.62) with respect to d_i $(i = 1 \ldots p)$ subject to the constraints

$$d_i = 0 \qquad i > r+1$$

$$d_1 = d_2 = \ldots = d_r = 1$$

$$0 < d_{r+1} < 1 \ .$$

For a given rank r, the optimal weight will be

$$d^*_{r+1} = \frac{\tau^2_{r+1}}{1+\tau^2_{r+1}}$$

where

$$\tau_i^2 \underline{\underline{\text{def}}} \frac{\gamma_i^2 \lambda_i}{\sigma^2} \qquad i = 1 \ldots p. \tag{3.65}$$

The truncated principal component estimator with assigned rank r is the special case with constraints

$$\begin{cases} d_i = 0 & i > r \\ d_1 = d_2 = \ldots = d_r = 1 \end{cases}$$

which does not require minimisation for the selection d's.

(3) The ridge estimator $\tilde{\xi}_{ORE}$ can also be written in this general form with

$$d_i = \left(1 + \frac{k}{\lambda_i} \right)^{-1} \tag{3.66}$$

or

$$k = \frac{\lambda_i(1-d_i)}{d_i} \qquad i = 1 \ldots p \ . \tag{3.67}$$

where

k is a positive scalar constant.

Thus, (3.62) can be minimised with respect to d_i subject to the constraint that

$$\frac{\lambda_i(1-d_i)}{d_i} = \text{constant} \quad i = 1 \ldots p \ .$$

Unfortunately, this constrained minimisation does not yield a closed form solution for the d_i, hence no optimal k can be obtained explicitly. (Same result as in last section). However the optimal k can be approximated by a harmonic mean of the k_i from the optimal d_i's estimated unrestrictedly as follows.

(4) A global unrestricted minimisation of (3.62) with respect to d_i $(i = 1 \ldots p)$ gives the optimal weights

$$d_i^* = \frac{\tau_i^2}{1+\tau_i^2} \qquad i = 1 \ldots p$$

where τ_i^2 are defined in (3.65)

$$= \left[1 + \frac{\sigma^2}{\lambda_i \gamma_i^2} \right]^{-1}$$

that is

$$\tilde{\xi}_{GRE} = \begin{pmatrix} 1 + \frac{\sigma^2}{\lambda_1 \gamma_1^2} & & \bigcirc \\ & \ddots & \\ \bigcirc & & 1 + \frac{\sigma^2}{\lambda_p \gamma_p^2} \end{pmatrix}^{-1} \begin{pmatrix} \hat{\gamma}_1 \\ \\ \hat{\gamma}_i \\ \\ \hat{\gamma}_p \end{pmatrix} = (I+\Lambda^{-1}K)^{-1}\hat{\gamma} \qquad (3.68)$$

where $K = \text{diag}\{k_1 \ldots k_p\}$, $k_i = \frac{\sigma^2}{\gamma_i^2}$ $(i = 1 \ldots p)$ is a form of the Generalised Ridge Estimator with optimal k_i's.

It is easily seen from the Figure 3.1 that $\tilde{\xi}_{SJ}$ or $\tilde{\xi}_{SH}$ is some point on the line $\overline{OB}$ and $\tilde{\xi}_{ORE}$ is a point on a curve lying inside the triangle ΔOBA while $\tilde{\xi}_{GRE}$ can be anywhere within the rectangle with diagonal $\overline{OB}$

Note that by viewing the ridge estimators as sort of posterior mean with respect to a spherical normal prior located at the origin Leamer and

Chamberlain (1976) has shown that it can be represented as a matrix weighted average of all possible restricted estimates (in the space spanned by principal component points).

§3.6 The Bayesian Interpretation

The problem of multicollinearity is interpreted by the classicists as a lack of information from the data about some linear combination of the regression coefficients. A Bayesian is more fortunate in this situation as he can always supplement the lack of data information by a priori information. As we have noted, all the information for a Bayesian is summarised in his posterior distribution. Collinearity in the data implies that data information is diffuse and the posterior distribution approaches the prior distribution whereas if prior information is diffuse, it is well known that the posterior distribution coincides with the likelihood function. Formally, let the sample distribution and prior distribution of β be, respectively.

$$\hat{\beta} \sim N_p(\beta, \sigma^2(X'X)^{-1}) \tag{3.69}$$

$$\beta \sim N_p(\theta, \sigma_\beta^2 M^{-1}) \ . \tag{3.70}$$

Then the posterior distribution of β is

$$\beta | \hat{\beta} \sim N_p(\bar{\beta}, \Sigma) \tag{3.71}$$

where

$$\bar{\beta} = \left(\frac{1}{\sigma^2} X'X + \frac{1}{\sigma_\beta^2} M \right)^{-1} \left(\frac{1}{\sigma^2} X'y + \frac{1}{\sigma_\beta^2} M\theta \right) \tag{3.72}$$

and

$$\Sigma = \left(\frac{1}{\sigma^2} X'X + \frac{1}{\sigma_\beta^2} M \right)^{-1} . \tag{3.73}$$

Now, if there exist some vector c such that $c'\Sigma^{-1} = 0$, then (3.71)

will be improper since then Σ does not exist. This situation typically happens when the data is collinear and the prior is diffuse (i.e. $c'X' = 0$ and $\sigma_\beta^2 = \infty$). From (3.72) we see that

$$\left(\frac{1}{\sigma^2} X'X + \frac{1}{\sigma_\beta^2} M \right) \bar{\beta} = \frac{1}{\sigma^2} X'y + \frac{1}{\sigma_\beta^2} M\theta \ . \tag{3.74}$$

Premultiplication by c' with $c'X' = 0$ implies

$$\frac{1}{\sigma_\beta^2} c'M\bar{\beta} = \frac{1}{\sigma_\beta^2} c'M\theta$$

showing that the posterior mean $\bar{\beta}$ is the same as the prior mean θ. Similarly, one can show that the posterior variance is the same as the prior variance, hence the posterior distribution coincides with the prior distribution when the data is collinear. Without any prior information, the problem of collinearity is unsolvable. The usual solution for perfect collinearity by setting exact restrictions on certain linear combinations of the coefficients is equivalent to postulating that the prior distribution of these linear combinations is degenerate at the restricted value [see Zellner (1971)].

It is interesting to note that if the prior mean of β is 0 and $M = I_p$ in (3.70), the posterior mean in (3.71) has the form of an ordinary ridge estimator with k having an interpretation as the ratio of sample variance to prior variance. Hence, a small value of k implies having a vague prior whereas a large value of k implies a tight prior. Such prior will be useful if we have reason to believe that all elements of β are equally important and similar to one another - an assumption of exchangeability. If this assumption does not hold, ridge estimator may be a very poor alternative to least square estimate and other more reasonable prior is needed.

Leamer (1973) gave an excellent account of "harmful" collinearity

from a Bayesian point of view. He pointed out that the "ignorance" or reluctance to use prior information has prompted traditional users of least squares estimators to overlook the fact that sample evidence of one coefficient may be affected by that of other coefficients, especially so when the data is collinear. The habit of interpreting data evidence of one coefficient at a time as if the regressors were orthogonal has distorted the interpretation of actual multidimensional sample evidence. Given a prior distribution, the posterior distribution is fully defined and there is no ambiguity about estimates of the coefficient vector and thus no collinearity problem in the traditional sense. However, there remains the difficult problem in selecting an acceptable prior distribution.

When collinearity is present, the posterior distribution may be highly sensitive to changes in the prior. If prior information dominated sample evidence in all directions there would be no collinearity problem. If prior information is uncertain, then more careful selection of a prior distribution is needed when the likelihood function has peculiar shape due to collinearity.

Finally, the interpretation of ridge regression as a posterior mean through a conjugate normal prior has helped to enrich the class of ridge regression in the form of (3.72). It allows for the addition of any positive semidefinite matrix to $X'X$ as well as the pulling of OLS towards a more plausible location other than the origin. Analogous existence theorems can be obtained for these wider class of ridge estimators (see Haitovsky and Wax (1976)).

CHAPTER 4

POTENTIAL AND REALISED IMPROVEMENTS

§4.1 Introduction

The primary motivation for departure from the least square principle is that OLS may fail to provide stable and precise estimates when used to analyse non-experimentally generated data. Also, the fact that LS estimators are indeed inadmissible when there are more than two parameters to be estimated has stimulated enthusiastic search for an adequate and better alternative. Theoretically, under an invariant loss function, a minimax alternative to OLS always exists when the dimension of the parameter space exceeds 2 ($p > 2$). The gain over OLS is potentially greater when the signal to noise ratio (or non-centrality parameter) is smaller. Since the invariant loss function is not the only one commonly employed by econometricians, the potential gain may not be realised when the minimax alternatives are used under some other quadratic loss functions. Specifically, in multiple regression, the collinearity structure of the regressors plays a crucial role as to whether the minimax alternatives are very much, if at all, superior to the least squares estimator. The introduction of ridge regression was designed to exploit the collinearity structure of the regressors. Despite the promising theoretical results of the "existence" theorems, none of the commonly known ridge estimators are minimax.[1] That means that the ridge estimator (based on a non-stochastic choice of the ridge parameter) does have dramatic gain over

---

[1] Strawderman (1978) developed a minimax adaptive generalised ridge regression estimator which depends on the loss function considered. He pointed out that the ordinary ridge regression estimator is optimal under the quadratic loss function L_W with $W = (X'X)^{-2}$.

the least squares estimator in a particular region of the parameter space but also loses miserably outside this region. Since the optimal ridge parameter, which is a function of the unknown true parameters, is unobservable, an estimate must be used. The result is an adaptive ridge estimator having intractable distributional properties. Comparisons among different versions of adaptive ridge estimators must invariably rely on simulation studies. The difficulty in specifying representative experimental designs has generated a vast number of conflicting Monte Carlo results. In this chapter, we shall compare various versions of adaptive ridge estimators that seem to survive different simulations in the literature. We study them in a unified Monte Carlo experiment with a view to screening out the best performer. The proposed experimental design of Baldwin and Hoerl (1978) was used in the Monte Carlo study. Finally, some 'good' ridge estimators are applied in estimating the investment function based on the structural equation specified by Christ (1966) to obtain a feel of how the ridge estimator performs in a real data situation.

§4.2 Adaptive Ridge Estimators

We consider the usual regression equation

$$y^* = \underset{\sim}{\ell}\alpha_o + W\alpha + \underset{\sim}{u} \qquad (4.1)$$

where y^* is $T\times 1$ vector of observations on the dependent variable, $\underset{\sim}{\ell}$ is a $T\times 1$ vector of unit elements, W is a $T\times p$ matrix of observations on p non-dummy regressors. We adopt the *convention* of standardising the regressors so that

$$y = X\beta + \varepsilon \qquad (4.2)$$

where

$$y = Ay^*$$

$$A = I - \frac{1}{T}\underset{\sim}{\ell}\underset{\sim}{\ell}'$$

$$X = AWD^{\frac{1}{2}}$$

$$D^{-1} = \text{diag}\left\{ \sum_{i=1}^{T} (w_{1i}-\bar{w}_1)^2, \ldots, \sum_{i=1}^{T} (w_{pi}-\bar{w}_p)^2 \right\}$$

$$\beta = D^{-\frac{1}{2}}\alpha$$

so that $X'X = D^{\frac{1}{2}}W'AWD^{\frac{1}{2}}$ is a correlation matrix of the regressors. We are interested in obtaining a ridge estimate of the elements of α. Admittedly, the indirect ridge estimate for α implied by the ridge estimate for β will be in general different from the direct ridge estimate for α. Also, the loss function used for β will then not be the same as the loss function[2] used for α. However, the reason for adopting such a convention is to ameliorate the numerical solution of the normal equations and to avoid the problem of numerical domination by certain dimensions because of different unit of measurements in economic variables. A more convincing justification for adopting the convention was given by Vinod (1978) who observed that the re-parameterised model is more likely to satisfy the exchangeability assumption, if ridge regression is interpreted from a Bayesian point of view. The canonical representation of model (4.2) is, continuing with the notations in §3.2,

$$y = XGG'\beta + \varepsilon \qquad (4.3)$$

where

$G'\beta = \gamma$ are the true principal components.

The least squares estimator of the true principal components is given by

$$\hat{\gamma} = G'\hat{\beta} = \Lambda^{-1}G'X'y \qquad (4.4)$$

---

[2] e.g. $L_W(\hat{\alpha}) = (\hat{\alpha}-\alpha)'W(\hat{\alpha}-\alpha) = (\hat{\beta}-\beta)'D^{\frac{1}{2}}WD^{\frac{1}{2}}(\hat{\beta}-\beta) \neq L_W(\hat{\beta})$.

with mean and variance given respectively by

$$\mathcal{E}\hat{\gamma} = G'\beta = \gamma \tag{4.5}$$

and

$$\text{var } \hat{\gamma} = \sigma^2 G'(X'X)^{-1}G = \sigma^2 \Lambda^{-1} \ . \tag{4.6}$$

The coefficient of determination of OLS regression is

$$R^2 = \frac{\hat{\beta}'X'y}{y'y} = \frac{r'r}{y'y} \tag{4.7}$$

where

$$X = P\Lambda^{\frac{1}{2}}G' \quad \text{and} \quad r = P'y = (r_{y1} \ \ldots \ r_{yp})'$$

the p columns of $(T\times 1)$ vectors of P are principal co-ordinates of X such that $P'P = I_p$ and t-ratios for testing $\gamma_i = 0$ $(i = 1 \ \ldots \ p)$ are

$$t_i = \frac{\hat{\gamma}_i}{\sqrt{\frac{y'y(1-R^2)}{(T-p-1)\lambda_i}}} = \frac{r_{yi}}{\sqrt{\frac{y'y(1-R^2)}{T-p-1}}} \qquad i = 1 \ \ldots \ p \tag{4.8}$$

which is seen to be independent of λ_i.

Consider now the general shrinkage estimator of

$$\tilde{\beta}_{GSH} = G\Delta\hat{\gamma} = G\tilde{\xi} \tag{4.9}$$

where

$\Delta = \text{diag}\{\delta_1 \ldots \delta_p\}$, $0 \leq \delta_i \leq 1$ $\quad i = 1 \ \ldots \ p$

$\tilde{\xi}$ is given in (3.60) and $\delta_i = \sum_{j=i}^{p} a_j = d_i$ as described in section 3.5.

The mean and variance of $\tilde{\beta}_{GSH}$ are respectively

$$\mathcal{E}\tilde{\beta}_{GSH} = G\Delta\gamma \neq \beta \quad \text{unless} \quad \Delta = I_p \tag{4.10}$$

and

$$\text{Var } \tilde{\beta}_{GSH} = \sigma^2 G\Delta\Lambda^{-1}\Delta G' = \sigma^2 G\Delta^2\Lambda^{-1}G' \ . \qquad (4.11)$$

From (4.11), it is clear that given a value of σ^2, the variance of $\tilde{\beta}_{GSH}$ would be reduced if δ_i is taken to be small when λ_i is small. This suggests the advantage of having a "declining deltas" scheme for ridge analysis. In other words, it would achieve a bigger reduction in variance if the shrinkage factor $(\delta_i, i = 1 \ldots p)$ are chosen such that the components corresponding to the smaller e-roots will be shrunken more to zero. This is intuitively sensible because the components corresponding to small e-roots are those which are imprecisely estimated by OLS.

The objective of ridge analysis is to choose a set of δ's such that $\tilde{\beta}_{GSH}$ has desirable properties.[3] As was established in Chapter 3, the Generalised Ridge Estimator (GRE) is a member of the class of general shrinkage estimator in (4.9) with the special choice of deltas

$$\delta_i = \frac{\lambda_i}{\lambda_i + k_i} \qquad i = 1 \ldots p \qquad (4.12)$$

where k_i $(i = 1 \ldots p)$ are the ridge parameters to be chosen. It is well-known that the total mean square error (TMSE) (see Lawless (1978))

$$\text{TMSE} = \frac{1}{\sigma^2} R_I(\tilde{\beta}_{GRE}) = \sum_{i=1}^{p} \left\{ \frac{\lambda_i}{(\lambda_i+k_i)^2} + \frac{1}{\sigma^2}\left(\frac{k_i\gamma_i}{(\lambda_i+k_i)} \right)^2 \right\} \qquad (4.13)$$

will be minimised when we set $k_i^* = \dfrac{\sigma^2}{\gamma_i^2}$ $(i = 1 \ldots p)$. So that, the "optimal" ridge estimator is

$$\tilde{\beta}_{OPT} = G\Delta^*\hat{\gamma} \qquad (4.14)$$

---

[3] Namely, estimates become stable and with smaller total mean square error than LS estimates, hopefully, wrongly signed LS estimates will be corrected to have the a-priori expected sign.

where $\Delta^* = \text{diag}\{\delta_1^* \ldots \delta_p^*\}$ and $\delta_i^* = \frac{\lambda_i}{\lambda_i+k_i^*} = \frac{\gamma_i^2\lambda_i}{\gamma_i^2\lambda_i+\sigma^2}$. Unfortunately, the "optimal" ridge estimator is not feasible as the ridge parameters are functions of unknown quantities. One way to get a feasible estimator is simply to replace the unknowns by LS estimates. Hence, we have the adaptive version of the 'optimal' estimator, or the "naive" estimator[4] as referred to by Guilkey and Murphy (1975) and Lawless and Wang (1976).

$$\tilde{\beta}_{NAIVE} = G\hat{\Delta}\hat{\gamma} \tag{4.15}$$

where

$$\hat{\Delta} = \text{diag}\{\hat{\delta}_1 \ldots \hat{\delta}_p\} \quad \text{and} \quad \hat{\delta}_i = \frac{\lambda_i}{\lambda_i+\hat{k}_i}$$

$$\hat{k}_i = \frac{\hat{\sigma}^2}{\hat{\gamma}_i^2} \quad (i = 1 \ldots p).$$

Note that, this involves estimation of p more parameters $(k_1 \ldots k_p)$ in addition to the $p+1$ unknown parameters $\{\gamma_1 \ldots \gamma_p, \sigma^2\}$. Obviously, this is a significant drain in degrees of freedom especially when the number of observations is small. On the other hand, the λ_i's so obtained may not be declining according to λ_i's and it is possible that $\tilde{\beta}_{NAIVE}$ does not reduce the variance substantially enough. Although reduction in variance does not necessarily guarantee a reduction in mean square error as the bias term has still to be taken into account, yet the expression $\delta_i^* = \frac{\gamma_i^2\lambda_i}{\gamma_i^2\lambda_i+\sigma^2}$ suggests that when $\lambda_i > \lambda_j$, one expects that $\gamma_i^2\lambda_i > \gamma_j^2\lambda_j$ unless there is strong evidence that γ_i^2 is much smaller than γ_j^2. However, even when such evidence exists, one can still restrict

---

[4] Compare this with the Huntsberger (1955)'s combined estimators in Chapter 2.

attention to $\delta_i > \delta_j$ simply because relatively less can be gained by introducing more bias into $\hat{\gamma}_i$ than into $\hat{\gamma}_j$, the latter being more imprecise because of the ordering of their variances $\frac{\sigma^2}{\lambda_j} > \frac{\sigma^2}{\lambda_i}$.

Building on this, Obenchain (1975) developed a likelihood selection criterion for estimating the ridge parameters. In order to save degrees of freedom, he considered that the k_i's can be related to only two unknown parameters (q,k), say, such that

$$k_i = k\lambda_i^q \qquad i = 1 \ldots p \; . \tag{4.16}$$

This is equivalent to imposing $p-2$ restrictions on a model with $2p+1$ parameters $\{\sigma^2, \beta_1 \ldots \beta_p, k_1 \ldots k_p\}$. Hoerl and Kennard's (1970) ordinary ridge family has $q = 0$, which in fact imposes $p - 1$ restrictions. The likelihood selection criterion developed by Obenchain (1975) is to test whether such restrictions are acceptable based on the assumption that the distribution of the residual errors is normal.

Specifically, in order to estimate the p unknown "optimal" ridge parameters unrestrictedly, we can write, from (4.14),

$$\gamma_i^2 = \frac{\delta_i^* \sigma^2}{\lambda_i (1-\delta_i^*)} \qquad i = 1 \ldots p \tag{4.17}$$

and the log likelihood function as

$$\ell n \, L(\gamma, \sigma^2) = -\frac{T}{2} \ell n(2\pi\sigma^2) - \frac{1}{2\sigma^2} (y'y - 2y'P\Lambda^{\frac{1}{2}}\gamma + \gamma'\Lambda\gamma) \; . \tag{4.18}$$

If $\tilde{\sigma}^2$ is an estimator of σ^2, the corresponding maximum likelihood estimator of γ, subject to (4.17), can be shown to be

$$\tilde{\gamma} = \tilde{\sigma}^2 \Lambda^{-\frac{1}{2}} \psi \tag{4.19}$$

where ψ is $(p \times 1)$ vector whose elements are

$$\psi_i = \sqrt{\frac{\delta_i^*}{1-\delta_i^*}}\ \text{sign}(r_{yi}) \qquad i = 1 \ldots p \ . \tag{4.20}$$

So that we have a concentrated likelihood function of $\tilde{\sigma}$, which is then maximised when $\frac{1}{\tilde{\sigma}} \left\{ -T - \left(\frac{1}{\tilde{\sigma}}\right)\sqrt{y'y}\,\psi' r + \left(\frac{1}{\tilde{\sigma}}\right)^2 y'y \right\} = 0$

$$\text{i.e.} \qquad \tilde{\sigma} = \frac{2\sqrt{y'y}}{\sqrt{(r'\psi)^2+4T}+r'\psi} \tag{4.21}$$

(the positive root of the quadratic equation in $\left(\frac{1}{\tilde{\sigma}}\right)$ is taken).
The corresponding conditional maximum likelihood estimator of $\beta = G\gamma$ is

$$\tilde{\beta} = \tilde{\sigma} G \Lambda^{-\frac{1}{2}} \psi \ \neq\ \tilde{\beta}_{GRE} = G\Delta^* \hat{\gamma} \tag{4.22}$$

and the minimum of $-2\,\ell n\, L(\tilde{\sigma})$, after substitution of (4.21) and (4.22) into (4.18),

$$-2\,\ell n\, L(\tilde{\sigma}) = T\,\ell n\,(2\pi\tilde{\sigma}^2) + \psi'\psi - \frac{r'\psi\sqrt{y'y}}{\tilde{\sigma}} + \frac{y'y}{\tilde{\sigma}^2} \ . \tag{4.23}$$

Note that this is *still conditional* on the unknown ψ (since δ_i^* and k_i^* are unknown). If attention is restricted to the Hoerl and Kennard's 1-parameter family ($q = 0$), then Δ^*, ξ, $\tilde{\sigma}$, $\tilde{\beta}$ and $-2\,\ell n\, L(\tilde{\sigma})$ can be computed for all values of the single ridge parameter k in (4.16). The Δ^* at which $-2\,\ell n\,\tilde{L}$ is minimised by the choice of K^+, denoted by Δ^+, leads to the "restricted" estimator of $\tilde{\beta}_{OPT}$ at (4.14) as

$$\tilde{\beta}^+_{OPT} = G\Delta^+\hat{\gamma} \ . \tag{4.24}$$

This $\tilde{\beta}^+_{OPT}$ results from having estimated only one more parameter from the data than are used in $\hat{\beta}$.

With no restrictions on the p parameters k_i ($i = 1 \ldots p$), the global minimum of $-2\,\ell n\,\hat{L}$ is achieved at

$$\hat{\Delta} = \text{diag}\{\hat{\delta}_1 \ \ldots \ \hat{\delta}_p\} \tag{4.25}$$

where

$$\hat{\delta}_i = \frac{\lambda_i}{\lambda_i + \hat{k}_i^*} = \frac{1}{1 + \frac{\hat{\sigma}^2}{\lambda_i \hat{\gamma}_i^2}} \quad \text{and} \quad \hat{\sigma}^2 = \frac{(1-R^2)y'y}{T}$$

which shows that the "naive" estimator at (4.15) is in fact the unrestricted maximum likelihood estimator of $\tilde{\beta}_{OPT}$ at (4.14).

The validity of restricting attention to the 1-parameter "ordinary ridge" family of Hoerl and Kennard can now be tested using the likelihood ratio statistic.

$$-2 \ell n \left(\frac{\tilde{L}}{\hat{L}} \right) \sim \chi^2_{p-1} \tag{4.26}$$

where $-2 \ell n \hat{L} = T \ell n (2\pi e \hat{\sigma}^2)$ is the unrestricted minimum of $-2 \ell n L$. In the process of minimising $-2 \ell n \tilde{L}$, we can identify the ridge parameter k^+ giving an adaptive ordinary ridge estimator, $\tilde{\beta}^+_{OPT}$, which will be studied in the subsequent Monte Carlo experiment and applied to the analysis of the investment data of Christ (1966).

Obenchain (1975) also proposed a preliminary test of whether a simple uniform shrinkage scheme would be suitable before one introduces a declining shrinkage scheme. Since there exists an uniform shrinkage minimax alternative to OLS when the design matrix is orthogonal, the test is tantamount to asking if the collinearity structure of the design matrix is "harmful" or not. A non-parametric test based on ranks is the Positive Correlation Between Spread $\sqrt{\lambda_i}$ and the Absolute values of r_{yi}, abbreviated (PCSA). It was based on the assumption that $|r_{yi}|$ would adequately reflect the absolute value of the true coefficients $|\gamma_i|$. As pointed out earlier, if $\gamma_i^2 \lambda_i > \gamma_j^2 \lambda_j$, then the optimal deltas have the property that $\delta_i^* > \delta_j^*$. Note that such a test could not be very

powerful if the number of parameters p is less than 10.

Hoerl and Kennard (1970) relied on a graphical ridge trace to identify the region of ridge parameters where the ridge estimates begin to stabilise. Such a choice of ridge parameter depends on the subjective judgement of the investigator and hence lacks uniqueness. Obenchain's proposal of SSCBC (see §3.4) may help to quantify the region and identify the point at which SSCBC is minimum as the ridge parameter. Although there may be more than one local minimum of SSCBC along the ridge trace, in practice this method often leads to a unique choice of k. We shall denote the ridge estimator corresponding to such a choice of ridge parameter by $\tilde{\beta}_{SSCBC}$.

One other criterion for selecting the ridge parameter along a ridge trace is the Mallow's C_p statistic as discussed in §3.4. This ridge estimate denoted by $\tilde{\beta}_{CP}$ will be compared to the above two in our illustrative example.

Other adaptive ridge estimators which are not obtained by means of the ridge trace technique (hence involves less computation) include the estimates of k by Hoerl, Kennard and Baldwin (1975) and that by Lawless and Wang (1976). These estimators, denoted respectively by $\tilde{\beta}_{HKB}$ and $\tilde{\beta}_{LW}$, have the advantage of being simple and operational. Both of them have a Bayesian justification. Specifically, assuming a spherical normal prior distribution for β with mean 0 and variance σ_β^2 the ridge parameter k is equivalent to the variance ratio $\frac{\sigma^2}{\sigma_\beta^2}$. Hoerl, Kennard and Baldwin estimated σ^2 by $\hat{\sigma}^2$ and σ_β^2 by $\frac{\hat{\beta}'\hat{\beta}}{p}$ so that $k_{HKB} = \frac{p\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}}$. They also justified it as the harmonic mean of $\frac{\hat{\sigma}^2}{\hat{\gamma}_i^2}$ $(i = 1 \ldots p)$. As we know, $\frac{\hat{\beta}'\hat{\beta}}{p}$ is not an ideal estimate of σ_β^2. In fact, the marginal distribution of $\hat{\beta}$ is

$$\hat{\beta} \sim N(0, \sigma^2(X'X)^{-1} + \sigma_\beta^2 I_p) \ . \tag{4.27}$$

Consider now the quantity,

$$\mathcal{E}\hat{\beta}'X'X\hat{\beta} = \text{tr}\, X'X\, \mathcal{E}\hat{\beta}\hat{\beta}'$$

from (4.27),

$$= \text{tr}(X'X)[\sigma^2(X'X)^{-1} + \sigma_\beta^2 I_p]$$

$$= \sigma^2 \text{tr}\, I_p + \sigma_\beta^2 \text{tr}\,(X'X) \ . \tag{4.28}$$

As X is standardised so that $X'X$ is a correlation matrix whose diagonal elements are units. Then an unbiased estimator for $\sigma^2 + \sigma_\beta^2$ is

$$\frac{\hat{\beta}'X'X\hat{\beta}}{p} \ . \tag{4.29}$$

Alternatively, if σ^2 is known, an unbiased estimator of $\frac{\sigma_\beta^2}{\sigma^2}$ is

$$\frac{\hat{\beta}'X'X\hat{\beta}}{p\sigma^2} - 1 \ . \tag{4.30}$$

Lawless and Wang (1975) argued that σ_β^2 will be much larger than σ^2 and proposed that

$$k_{LW} = \frac{p\sigma^2}{\hat{\beta}'X'X\hat{\beta}} \tag{4.31}$$

to be a reasonable estimate of $\frac{\sigma^2}{\sigma_\beta^2}$. Note that (4.31) can be represented in terms of the coefficient of determination of the least squares regression because $R^2 = \frac{\hat{\beta}'X'X\hat{\beta}}{y'y}$ and $\hat{\sigma}^2 = \frac{y'y(1-R^2)}{T-p-1}$. Thus, an equivalent estimate of $\frac{\sigma^2}{\sigma_\beta^2}$ is

$$k_{LW} = \frac{p(1-R^2)}{(T-p-1)R^2} \tag{4.32}$$

and is a particularly convenient formula to program.

The Monte Carlo experiment to follow also considers other more complicated estimators of k, namely, McDonald and Galarneau (1975)'s "correct length" criterion and Swamy, Mehta and Rappoport (1978)'s "strong condition" criterion. Both of them require solutions of implicit functions in k and are more troublesome computationally. Such efforts could only be justified if there is strong evidence of improvement in their favour. They were denoted by $\tilde{\beta}_{MG}$ and $\tilde{\beta}_{SMR}$ respectively. Note that we have tried the two versions of "strong condition" choice of k suggested in Swamy et al (1978) but we only present the result of the second version as it is uniformly better.

§4.3 Design of Monte Carlo Experiments

In many of the Monte Carlo experiments studied by independent investigators, the data structure is usually chosen from a real life regression problem. For a given set of true regression coefficients, random numbers are generated to represent "typical" regression problems. In what follows, the more recent and accessible published results will be discussed. Many other studies are referred to in Vinod's (1978) survey.

Newhouse and Oman (1971), McDonald and Galarneau (1975), Hoerl and Kennard (1976), and Hoerl, Kennard and Baldwin (1975) examined OLS and some versions of the adaptive ridge estimators. Lawless and Wang (1976) and Lawless (1978) compared various versions of adaptive ridge estimators to the principal components estimator and the Stein positive part estimators respectively. A comprehensive study of Dempster, Schatzoff and Wermuth (1977) investigated 57 different estimators' (basically grouped as OLS, ridge-type estimators, shrunken estimators, and variable selection procedures) performance on 160 model configurations. Gunst and Mason (1977) reported experimental comparisons of Latent Root Regressions, principal component estimators with a version of the adaptive

ridge estimator while Wichern and Churchill (1978) reported the simulation results of various versions of adaptive estimators including a ridge-trace estimator identified by Vinod's (1976) Index of Stability of Relative Magnitude (ISRM). Finally, Swamy, Mehta and Rappoport (1978) reported the good performance of their choice of ridge parameter and of the Hoerl-Kennard-Baldwin choice relative to the "correct-length" choice of McDonald and Galarneau (1975). Vinod (1978) studied the performance of a positive-part version of the Bhattacharya estimator and the Stein estimator relative to that of OLS and $\tilde{\beta}_{HKB}$. Almost all of the studies have reported the superiority of ridge estimators over OLS, although there is wide disagreement about the "optimum" ridge estimator. Also, there has emerged a consensus that the performance of the biased estimators relative to the OLS is dependent on one or more of the following factors:

(i) the orientation of the true coefficient vector, β, with respect to the e-vectors of $X'X$. Theorem 3.3.3 has established that the straight ridge estimator will be least biased if β lies on the direction of the e-vector corresponding the largest e-root and most biased if β lies in the direction of the e-vector corresponding to the smallest e-root

(ii) the magnitude of β relative to σ (i.e., the signal-noise ratio)

(iii) the conditioning of the data (i.e. the degree of multi-collinearity reflected by the set of e-roots)

(iv) the number (p) of regression coefficients to be estimated (dimensionality).

The desirability of a broad comparison of various adaptive ridge estimator to cover a wide range of dimensionality and conditioning was stressed by Hoerl in his comment on the paper of Dempster et al (1977). He suggested the need of a standardised design that typifies the regression problem so as to avoid unnecessary repetition of simulation. To this effect, Baldwin and Hoerl (1978) presented simulation model designs which, given the data conditioning and dimensionality, cover a wide range of signal to noise ratios. We would, therefore, take advantage of two of those designs to carry out a Monte Carlo study on various versions of adaptive estimators and other commonly used biased estimators in econometrics. Typically, we have used and modified their designs as given in Table 4.3.1 and Table 4.3.2 which are for $p = 4$ and $p = 10$ respectively, since we believe they represent the sort of model size and conditioning for standardised data sets commonly encountered in econometrics. Observing that σ^2 has only the effect of scaling TMSE, we select $\sigma^2 = 1$ for convenience so that the signal-noise-ratio (r^2) is the same as the squared distance of the true coefficient vector. The range of values of signal-noise ratio considered are $(1, 10, 100, 10^3, 10^4, 10^5)$. For each signal-noise ratio, (or vector length in our case), there are at least three different orientations of the true coefficient vector, viz., corresponding to the least, intermediate and the most biased direction for the straight ridge estimators. In all, there are twenty two different specifications for a design of given size (p) and a given conditioning.

For each specification, OLS estimates are replicated as

$$\hat{\gamma}^{(m)} \sim N_p(\gamma^{(m)}, \Lambda^{-1}) \qquad p = 4 \text{ or } 10 \qquad m = 1,\ldots,22 \tag{4.33}$$

1,000 replications are generated for models with $p = 4$, but only 100 replications are generated for models with $p = 10$ in order to economise

Table 4.3.1

Designs for Model of Size p=4

e-roots are {2.2357, 1.5761, .18661, .00162}

$\lambda_1/\lambda_p = 1380.0617$

Model No.	r^2	γ_1^2	γ_2^2	... γ_4^2	$TMSE(\tilde{\beta}_{OPT})$	$TPSE(\tilde{\beta}_{OPT})$
1	1	1	0	0	.330	.737
2	1		.13	.13	.640	.815
2	1		0	1	.998	.002
4	10	10	0	0	.453	1.013
5	10	6	1.33	1.33	3.319	1.949
6	10	0	0	10	9.857	.016
7	100	100	0	0	.467	1.045
8	100	60	13.33	13.33	18.177	2.864
9	100	20	26.67	26.67	31.516	3.024
10	100	0	0	100	87.871	.142
11	10^3	10^3	0	0	.468	1.046
12	10^3	600	133.33	133.33	119.269	3.338
13	10^3	200	266.67	266.67	201.951	3.494
14	10^3	0	0	10^3	421.036	.682
15	10^4	10^4	0	0	.467	1.045
16	10^4	6,000	1,333.33	1,333.33	477.165	3.958
17	10^4	2,000	2,666.67	2,667.67	577.137	4.122
18	10^4	0	0	10^4	675.428	1.094
19	99999	99,999	0	0	.467	1.045
20	99999	59,999	13,333.33	13,333.33	693.845	4.312
21	99999	19,999	26,666.66	26,666.66	711.976	4.342
22	99999	0	0	99,999	719.155	1.165
					$TMSE(\hat{\beta})$ =623.724	$TPSE(\hat{\beta})$=5

Table 4.3.2

Designs for Model of Size p=10

e-roots are $\begin{Bmatrix} 4.250, & 1.500, & 1.250, & 1.000, & .778 \\ .600, & .400, & .200, & .020, & .002 \end{Bmatrix}$

$\lambda_1/\lambda_p = 2125.000$

Model No.	r^2	γ_1^2	γ_2^2 . . .	γ_{10}^2	$TMSE(\tilde{\beta}_{OPT})$	$TPSE(\tilde{\beta}_{OPT})$
1	1	1	0 . . .	0	.157	.665
2	1	.6	.044	.044	1.178	.847
3	1	0	0	1	.997	.002
4	10	10	0 . . .	0	.190	.807
5	10	6	.444	.444	3.603	2.779
6	10	0	0	10	9.764	.020
7	100	100	0 . . .	0	.196	.834
8	100	60	4.444	4.444	18.234	7.081
9	100	20	8.889	8.889	28.065	8.071
10	100	0	0	100	81.632	.163
11	10^3	10^3	0 . . .	0	.197	.839
12	10^3	600	44.444	44.444	78.798	9.511
13	10^3	400	66.667	66.667	102.286	9.754
14	10^3	0	0	10^3	312.213	.624
15	10^4	10^4	0 . . .	0	.198	.841
16	10^4	6,000	444.44	444.44	290.118	10.655
17	10^4	4,000	666.67	666.67	338.252	10.799
18	10^4	0	0	10^4	437.028	.874
19	99999	99,999	0 . . .	0	.198	.841
20	99999	59,999	4,444.44	4,444.44	487.558	11.189
21	99999	39,999	6,666.67	6,666.67	501.605	11.323
22	99999	0	0	99,999	456.124	.912
					$TMSE(\hat{\beta})$ =563.15	$TPSE(\hat{\beta})$=11

on computational time. Based on the OLS estimates, various biased estimators (see below) can be calculated for each replication. The total unweighted squared errors and total weighted squared errors of each biased estimator $\tilde{\beta}$ are averaged over all replications to provide estimates for Total Mean Squared Error (TMSE)

$$\text{TMSE}(\tilde{\beta}) = \frac{1}{\sigma^2} \mathcal{E}(\tilde{\beta}-\beta)'(\tilde{\beta}-\beta) = \frac{1}{\sigma^2} \mathcal{E}(\tilde{\gamma}-\gamma)'(\tilde{\gamma}-\gamma) \tag{4.34}$$

and Total Prediction Squared Error (TPSE)

$$\text{TPSE}(\tilde{\beta}) = \frac{1}{\sigma^2} \mathcal{E}(\tilde{\beta}-\beta)'X'X(\tilde{\beta}-\beta) = \frac{1}{\sigma^2} \mathcal{E}(\tilde{\gamma}-\gamma)\Lambda(\tilde{\gamma}-\gamma) \tag{4.35}$$

respectively.

As the true coefficient vector is known for each replication, we can also obtain the "optimal" ridge estimator $\tilde{\beta}_{OPT}$ which is not feasible in real data analysis. This estimator presumably minimises TMSE in the class of generalised ridge estimator and we expect that $\text{TMSE}(\tilde{\beta}_{OPT})$ and $\text{TPSE}(\tilde{\beta}_{OPT})$ to provide the approximate lower bound for the respective quantities. The two calculated values are listed in the last two columns of Table 4.3.1 and Table 4.3.2. Sampling errors excepted, they are lower than those of OLS (at the bottom of the respective columns) and the possibility of dramatic gain of ridge-type estimates are clearly evident for models of low signal-noise ratio irrespective of the dimensionality. Note that the potential improvement is greatest when the true coefficient vector is oriented in the direction of the e-vector corresponding to the largest e-root. Of course, this potential improvement may not be realised when an adaptive version of ridge estimator is used and it is the purpose of this Monte Carlo study to find out how much of the potential improvement will be realised.

§4.4 Outcome of Experiments

Based on the designs discussed in the preceding section, the following adaptive ridge estimators are considered.

$$\tilde{\gamma}_i^{(m)} = \left(\frac{\lambda_i}{\lambda_i + k_i} \right) \hat{\gamma}_i^{(m)} \qquad i = 1 \ldots p \qquad m = 1 \ldots 22 \tag{4.36}$$

(i) $k_i = \dfrac{\hat{\sigma}^2}{\hat{\gamma}_i^2}$ gives 'naive' ridge estimator, $\tilde{\gamma}_{NAIVE}$

(ii) $k_i = \dfrac{p\sigma^2}{\hat{\gamma}'\hat{\gamma}}$ gives the Hoerl-Kennard-Baldwin estimator, $\tilde{\gamma}_{HKB}$

(iii) $k_i = \dfrac{p\sigma^2}{\hat{\gamma}'\Lambda\hat{\gamma}}$ gives the Lawless-Wang estimator, $\tilde{\gamma}_{LW}$

(iv) $k_i = k_{MG}$ gives the McDonald-Galarneau estimator, $\tilde{\gamma}_{MG}$ so that $\tilde{\gamma}_{MG}'\tilde{\gamma}_{MG} = \hat{\gamma}'\hat{\gamma} - \hat{\sigma}^2 \sum_{i=1}^{p} \dfrac{1}{\lambda_i}$

(v) $k_i = p.k_{SMR}$ gives the Swamy-Metha-Rappoport type II estimator, $\tilde{\gamma}_{SMRII}$ where k_{SMR} is a solution of the equation

$$\sum_{i=1}^{p} \frac{k\hat{\gamma}_i^2\lambda_i}{(2\lambda_i + k)} = \sigma^2 .$$

We have also computed the truncated principal components estimates with the components corresponding to the two small e-roots being arbitrarily truncated (but the sum of retained e-roots accounts for at least 95% of the total sum). Therefore, the truncated principal components estimators for the models with $p = 4$ and $p = 10$ are denoted by $\tilde{\gamma}_{PC2}$ and $\tilde{\gamma}_{PC8}$ respectively. Finally, the Stein positive part estimator

Table 4.4.1

Estimated TMSE of Different Estimators for Twenty-two 4-Factor Models

(Based on 1,000 Replicates)

Theoretical TMSE for OLS = 623.724

{e-roots = 2.2357, 1.5761, 0.1866, 0.0016}

Models \ Estimator	PC2	STEIN+	NAIVE	HKB	LW	MG	SMRII
1	1.157	377.849	375.965	152.804	1.238	.411*	842.561
2	1.417	352.453	376.022	152.826	1.353*	80.827	842.562
3	2.157	303.274	376.037	153.212	1.874*	84.366	842.562
4	1.157*	618.448	376.027	153.557	3.004	139.110	9.940
5	3.817*	580.957	377.332	154.121	4.252	124.454	9.959
6	11.157	306.648	379.264	158.221	10.829*	63.080	842.562
7	1.157*	718.591	375.966	160.266	11.186	144.046	842.226
8	27.817	713.104	384.268	166.321	20.234*	96.427	7.090
9	54.497	700.017	390.025	172.804	31.648*	77.776	756.197
10	101.157	341.361	410.714	208.006	100.352	17.003*	842.562
11	1.157*	730.271	375.959	222.126	170.345	827.754	842.562
12	267.817	729.718	426.056	264.571	165.456	32.104*	842.562
13	534.497	728.377	467.463	309.994	200.448	6.863*	842.561
14	1001.157	617.465	632.702	601.811	994.377	103.589*	842.562
15	1.157*	731.448	375.957	506.017	594.486	842.511	842.562
16	2667.817	731.392	689.099	574.627	571.900	89.199*	842.562
17	5334.497	731.258	819.044	650.987	574.692*	670.894	842.562
18	10001.157	954.910	850.167	1270.202	9819.576	2818.962	842.562*
19	1.157*	731.565	375.957	697.773	715.826	842.562	842.562
20	26667.817	731.559	826.547	712.839	712.170*	2996.707	842.562
21	53334.476	731.546	779.356	728.680	711.505*	44206.095	842.562
22	100000	762.696	740.389*	822.949	87652.591	36120.65	842.562
No. of * scored	6	0	1	0	8	6	1

* = closest to minimum TMSE (attained by $\tilde{\gamma}_{OPT}$); underlined values indicate that the estimator has a smaller TMSE than OLS has with a probability greater than .9.

Note: It is possible that more than 10% of the replicates of a particular estimator have errors greater than that of OLS but the rest have very small errors [e.g. MG in Model 10].

FIGURE 4.4.1

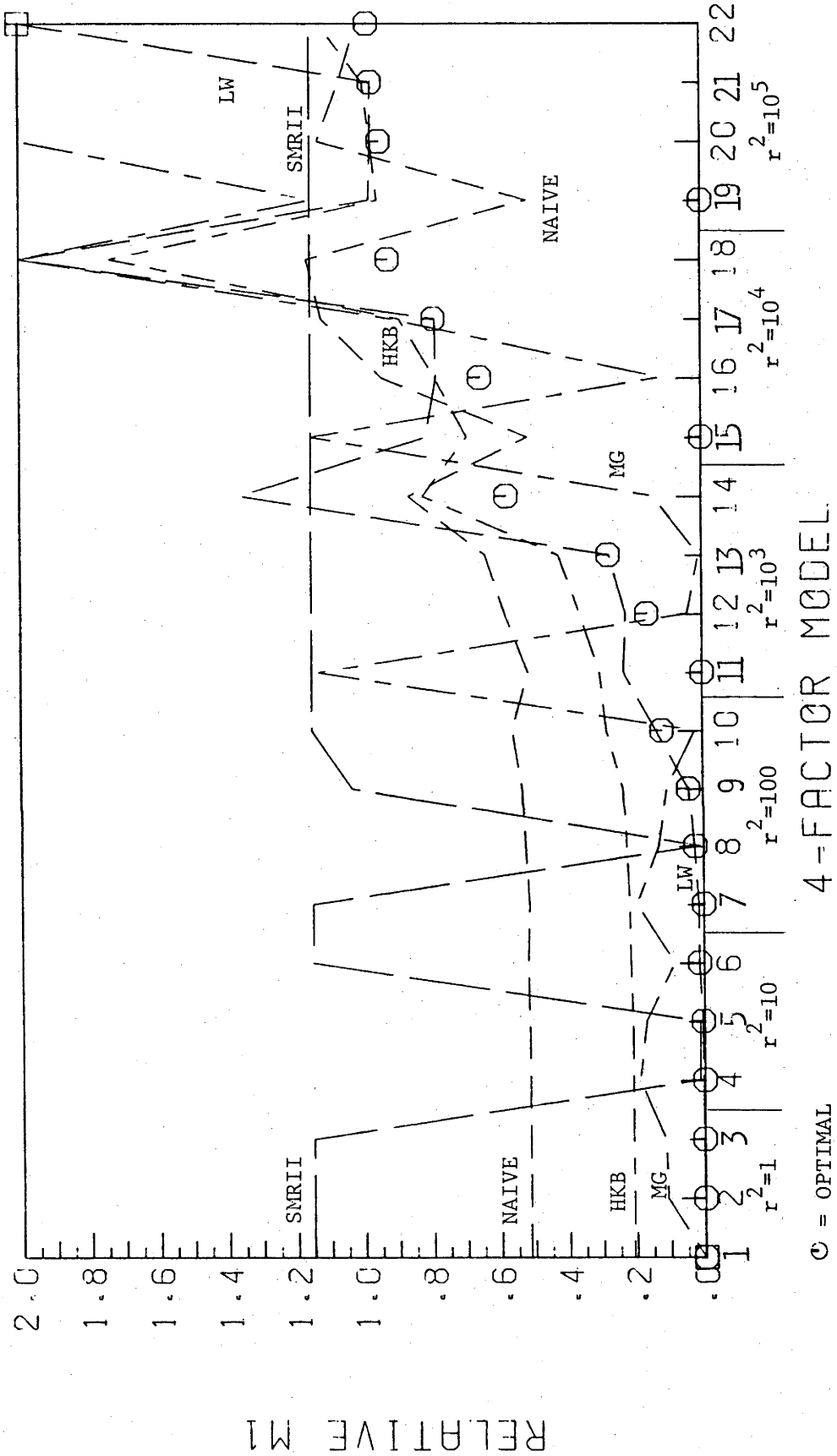

ESTIMATED TOTAL MSE
RELATIVE TO THAT OF OLS
RELATIVE M1
2.0
1.8
1.6
1.4
1.2
1.0
.8
.6
.4
.2
.0
SMRII
NAIVE
HKB
MG
LW
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
r²=1
r²=10
r²=100
r²=10³
r²=10⁴
r²=10⁵
4-FACTOR MODEL
Θ = OPTIMAL

Table 4.4.2

Estimated TPSE of Different Estimators for Twenty-Two 4-Factor Models

(Based on 1,000 Replicates)

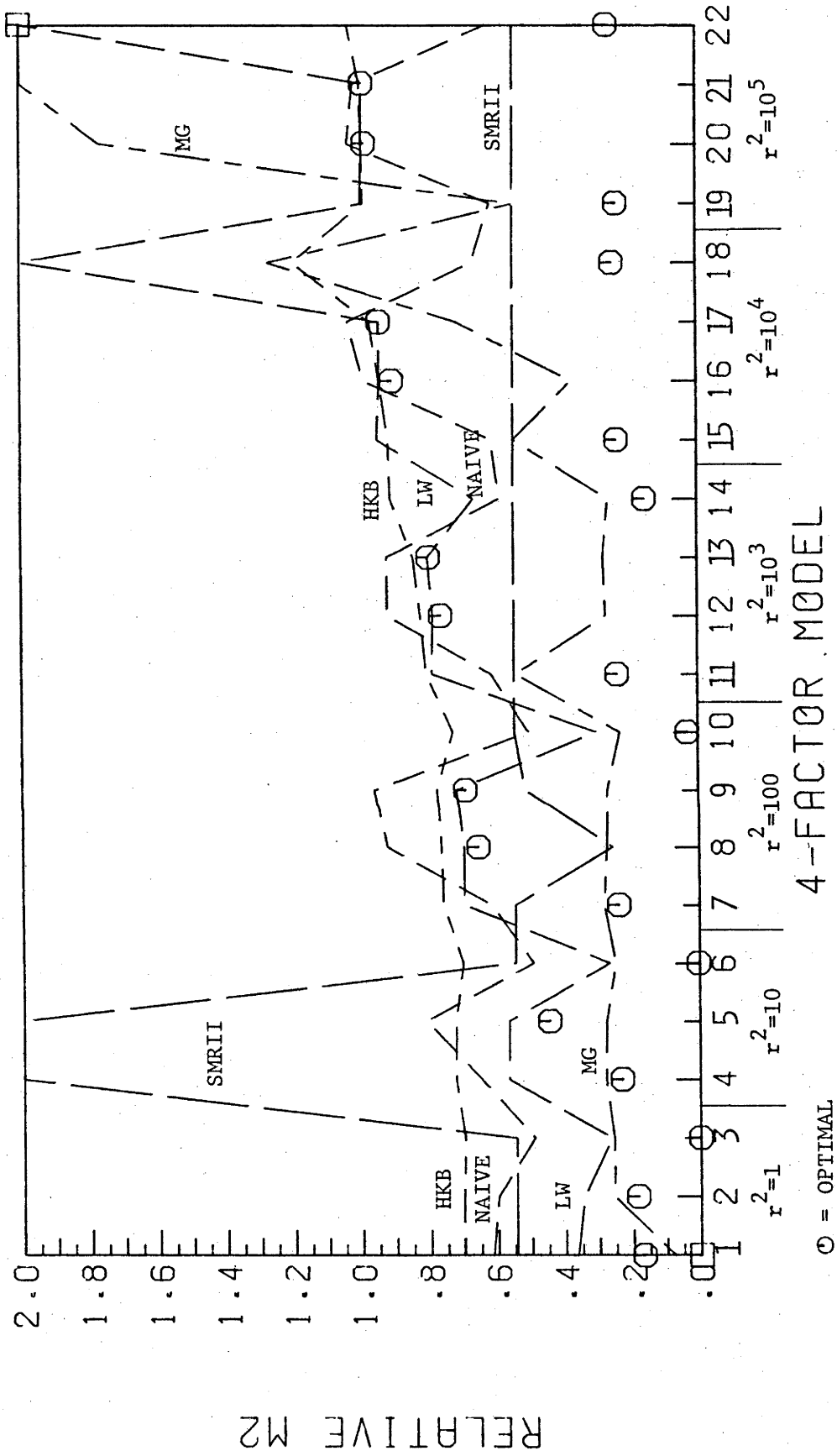
Theoretical TPSE for OLS = 5

e-roots - {2.2357, 1.5761, 0.1866, 0.0016}

Estimator / Models	PC2	STEIN+	NAIVE	HKB	LW	MG	SMRII
1	2.131	2.719	2.693	3.074	1.594	.337*	2.381
2	2.156	2.482	2.622	3.069	1.509	1.120*	2.381
3	2.133	1.734	2.145	3.050	1.148	1.122*	2.381
4	2.131	4.163	2.833	3.167	2.477	1.221*	22.224
5	2.382	4.067	3.546	3.160	2.471	1.210*	15.692
6	2.148	1.744	2.150	3.064	1.164	1.088*	2.381
7	2.131	4.360	2.695	3.321	3.045	1.235*	1.118
8	4.641	4.357	4.040	3.346	3.050	1.196*	1.118
9	7.152	4.342	4.202	3.390	3.139	1.181*	2.243
10	2.293	1.830	2.201	3.180	1.323	1.014*	2.381
11	2.131*	4.374	2.679	3.522	3.448	2.357	2.381
12	27.228	4.376	4.026	3.601	3.439	1.193*	2.381
13	52.327	4.377	4.029	3.691	3.506	1.223*	2.381
14	3.751	2.517	2.561	3.980	2.900	1.156*	2.381
15	2.131*	4.374	2.677	4.006	4.150	2.381	2.381
16	253.104	4.374	4.321	4.119	4.114	1.628*	2.381
17	504.079	4.376	4.525	4.245	4.120	3.116	2.381*
18	18.331	4.090	2.913	5.238	17.893	5.556	2.381*
19	2.131*	4.374	2.676	4.319	4.348	2.381	2.381
20	2511.864	4.374	4.529	4.344	4.342	7.735	2.381*
21	5021.597	4.374	4.452	4.369	4.341	7530.908	2.381*
22	164.130	4.346	2.735	4.521	144.900	59.505	2.381*
No. of * scored	3	0	0	0	0	14	5

* = closest to minimum TPSE (attained by $\tilde{\gamma}_{OPT}$); underlined values indicate that the estimator has a smaller TPSE than OLS has with a probability greater than .9.

FIGURE 4.4.2


ESTIMATED TOTAL PSE
RELATIVE TO THAT OF OLS
RELATIVE M2
2.0
1.8
1.6
1.4
1.2
1.0
.8
.6
.4
.2
.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
r^2=1
r^2=10
r^2=100
r^2=10^3
r^2=10^4
r^2=10^5
4-FACTOR MODEL
MG
SMRII
HKB
LW
NAIVE
SMRII
HKB
NAIVE
LW
MG
Θ = OPTIMAL

$$\tilde{\gamma}_{STEIN+} = \max \left[0,\ 1 - \frac{p-2}{\hat{\gamma}' \Lambda \hat{\gamma}} \right] \hat{\gamma} \qquad (4.37)$$

are also computed and compared.

Estimated TMSE and TPSE for the seven biased estimators for models with $p = 4$ are reported in Tables 4.4.1 and 4.4.2 respectively. The winner in each model among the seven competitors are marked by (*). From the bottom row of the tables, it can be seen that $\tilde{\gamma}_{LW}$ has won eight times out of twenty two models under TMSE criterion, while $\tilde{\gamma}_{MG}$ and $\tilde{\gamma}_{PC}$ are the two close runners up.

Under TPSE criterion, $\tilde{\gamma}_{MG}$ is the big winner scoring 14 points out of 22. This appears to agree with results of Wichern and Churchill (1978). However, considering the fact that $\tilde{\gamma}_{MG}$ requires more computation than $\tilde{\gamma}_{LW}$ does, and as the latter is not very much worse off in terms of TPSE, $\tilde{\gamma}_{LW}$ should be considered as a very strong competitor in view of its superior performance under the other loss function TMSE. To facilitate comparisons, the ratios of TMSE and TPSE for the five ridge estimators relative to those of OLS are graphed in Figures 4.4.1 and 4.4.2 respectively. Obviously, any curve below 1.0 implies that OLS has been beaten by that rule. The graphs reveal that $\tilde{\gamma}_{SMRII}$ is a highly unreliable choice and on many occasions can be more than twice as bad as OLS, while $\tilde{\gamma}_{LW}$ and $\tilde{\gamma}_{MG}$ are pretty stable for model of low signal-noise ratio. It is interesting to note that the minimax alternative, namely, $\tilde{\gamma}_{STEIN+}$ has not performed as well as other biased alternatives even under the invariant loss function (TPSE). This may be due to the fact that $\frac{2}{\lambda_p} = 1250$ is greater than $\text{tr}\,\Lambda^{-1} = 623.724$, violating the minimaxity condition for Stein rule to dominate OLS under TMSE. On the other hand, the behaviour of $\tilde{\gamma}_{PC2}$ is well expected. It depends heavily on the orientation of the true coefficient vector. Since information about the

components corresponding to the smallest e-root is truncated, if the true coefficient vector happens to lie in this direction, the loss will be great especially for large signal-noise ratio models. This completes the discussion of the results for models with $p = 4$.

Similar patterns of result emerge for models of size $p = 10$. $\tilde{\gamma}_{SMRII}$ remained as unreliable as before and $\tilde{\gamma}_{MG}$ has deteriorated. $\tilde{\gamma}_{LW}$ emerges as the winner under both criteria followed closely by $\tilde{\gamma}_{HKB}$. $\tilde{\gamma}_{PC8}$ has done very well under TMSE if the orientation of the true vector was in the direction of the e-vector corresponding to the largest e-root. $\tilde{\gamma}_{STEIN+}$ has performed well under the invariant loss function (TPSE) but the gain over OLS is only marginal. Nevertheless, such conservatism makes the Stein rule rather robust to the orientation of the true vector and it has never had a TPSE or TMSE exceeding those of OLS by a big margin. The results are conveniently summarised in Tables 4.4.3 and 4.4.4. The accompanying graphs in Figures 4.4.3 and 4.4.4 reveal graphically the performance of the five adaptive ridge estimators relative to that of OLS. When r^2 is smaller than 100, $\tilde{\gamma}_{LW}$ or $\tilde{\gamma}_{HKB}$ appear to be very attractive alternatives to OLS. Since there is usually prior infromation about the likely magnitude of the true coefficients of econometric models, the result of this Monte Carlo study provides some assurance that some versions of adaptive ridge estimator should be considered seriously as the gains over OLS could be more than halving the usual risks. Additionally, the ridge estimators have the ability to correct wrongly signed coefficients and to provide stabilised estimates which can have far-reaching economic implications if the coefficients are interpreted as elasticity parameters or as short run adjustment parameters. Some illustrations with real economic data seem worthwhile at this stage and we have chosen Christ's (1966) investment model and his data to do so.

Table 4.4.3

Estimated TMSE of Different Estimators for Twenty-two 10-factor Models

(Based on 100 Replications from Multivariate Normal Distributions)

Theoretical TMSE for OLS = 563.15

e-roots = $\begin{Bmatrix} 4.250, & 1.500, & 1.250, & 1.000, & 0.778 \\ 0.600, & 0.400, & 0.200, & 0.020, & 0.002 \end{Bmatrix}$

Models \ Estimator	PC8	STEIN+	NAIVE	HKB	LW	MG	SMRII
1	15.628	156.202	249.848	56.663	3.576*	176.655	866.061
2	16.362	134.450	250.265	57.090	4.149*	186.504	866.049
3	16.628	91.302	250.114	57.294	3.678	1.001*	866.141
4	15.628	389.986	249.803	57.300	8.396*	177.365	721.852
5	16.516	340.010	251.285	57.521	8.168*	182.070	29.486
6	25.628	97.056	253.821	63.894	12.646*	220.728	866.072
7	15.628*	515.199	249.793	62.600	30.188	184.580	794.383
8	24.517*	505.557	259.999	65.852	28.258	195.968	801.810
9	33.406	477.192	265.375	69.863	29.396*	196.565	810.016
10	115.628	149.362	287.423	128.773	102.241*	335.497	792.740
11	15.628*	532.372	249.793	101.361	161.094	853.948	866.360
12	104.517*	531.333	289.551	123.949	135.781	264.364	866.380
13	148.962	530.284	301.154	136.601	127.651*	255.856	866.380
14	1015.628	569.588	508.311*	654.782	997.193	1002.334	1000.098
15	15.628*	534.165	249.793	292.937	441.575	866.341	866.380
16	904.516	534.058	434.210	339.438*	412.786	619.294	866.380
17	1348.968	533.951	489.523	364.509*	394.133	2306.554	866.380
18	10015.628	1235.606	566.875*	1652.263	9874.776	5241.934	866.380
19	15.628*	534.346	249.793	491.721	523.620	866.380	866.380
20	8904.508	534.335	639.575	503.144*	518.995	2376.394	866.380
21	13348.97	534.32	627.10	509.436*	515.530	4118.581	866.380
22	100014.6	699.166	585.614*	796.295	92005.75	39174.11	866.380
No. of * scored	6	0	3	4	8	1	0

* = closest to minimum TMSE (attained by $\tilde{Y}_{OPT}$); underlined values indicate that the estimator has a smaller TMSE than OLS has with a probability greater than .9.

FIGURE 4.4.3

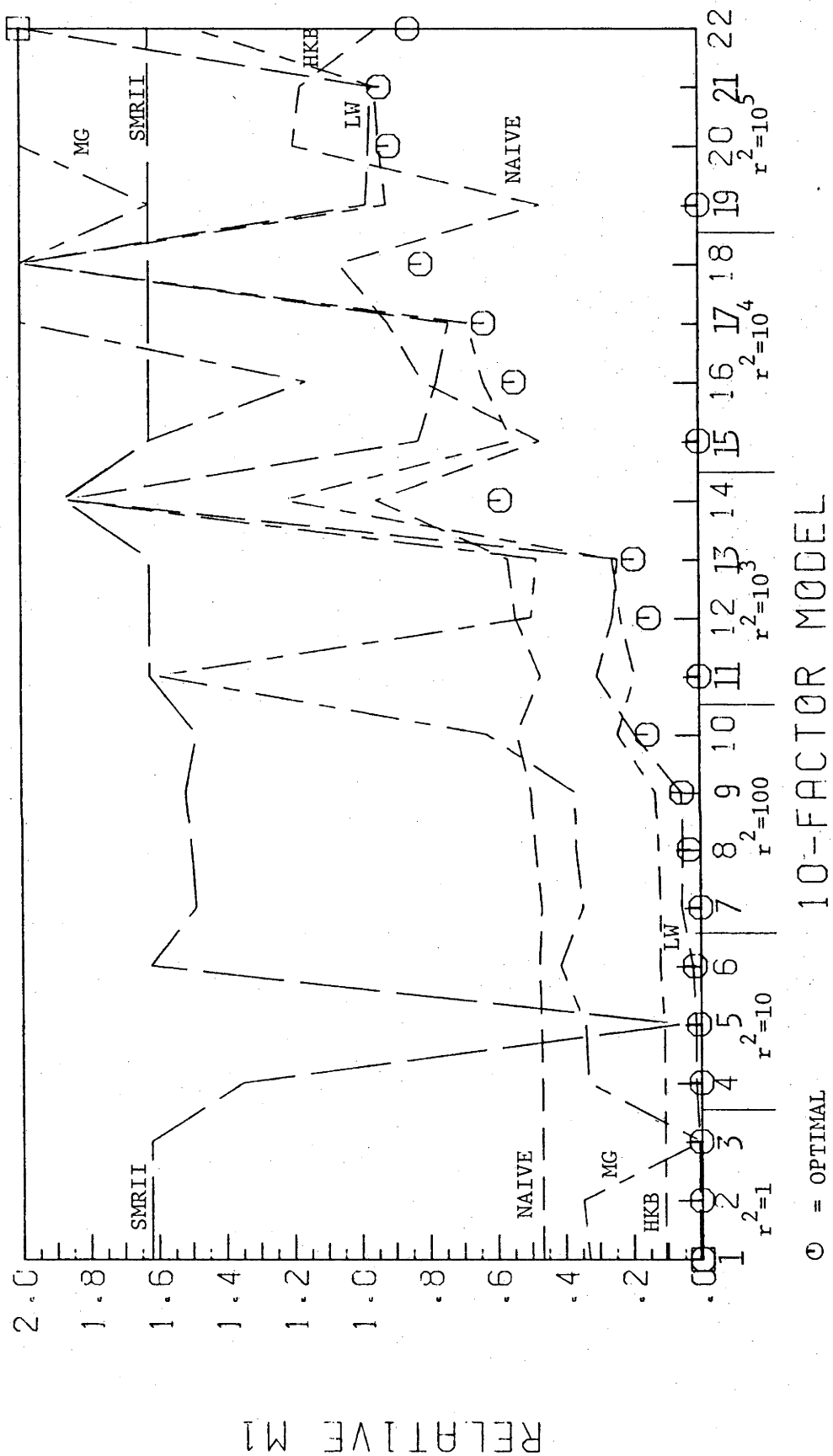

ESTIMATED TOTAL MSE
RELATIVE TO THAT OF OLS
RELATIVE M1
2.0
1.8
1.6
1.4
1.2
1.0
.8
.6
.4
.2
.0
SMRII
NAIVE
MG
HKB
LW
r^2=1
r^2=10
r^2=100
r^2=10^3
r^2=10^4
r^2=10^5
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Θ = OPTIMAL
10-FACTOR MODEL

Table 4.4.4

Estimated TPSE of Different Estimators for Twenty-Two 10-Factor Models

(Based on 100 Replications from Multivariate Normal Distribution)

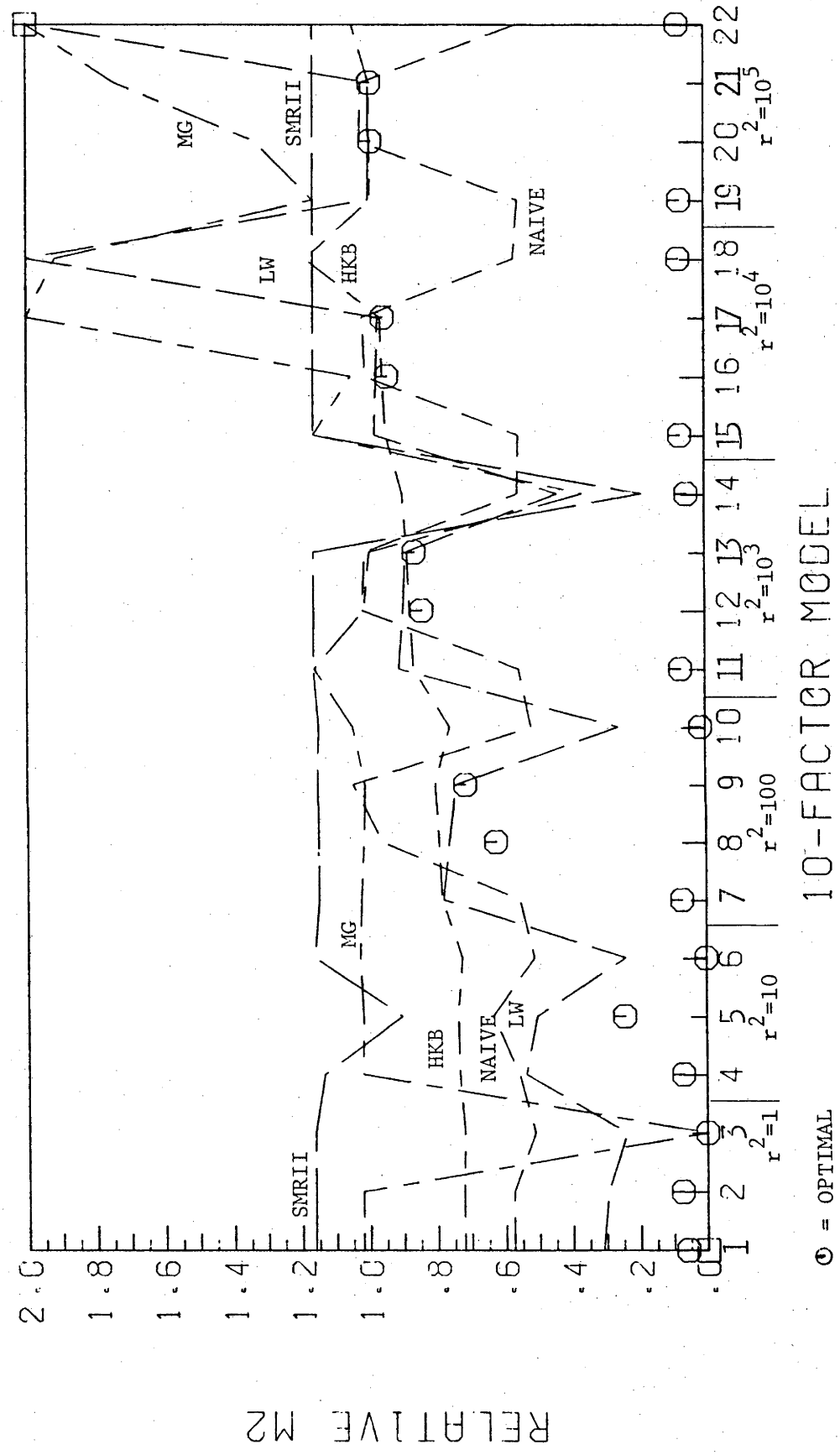
Theoretical TPSE for OLS = 11

$$\text{e-roots} = \begin{Bmatrix} 4.250, & 1.500, & 1.250, & 1.000, & 0.778 \\ 0.600, & 0.400, & 0.200, & 0.020, & 0.002 \end{Bmatrix}$$

Models \ Estimator	PC8	STEIN+	NAIVE	HKB	LW	MG	SMRII
1	9.186	4.796	6.508	8.158	3.510*	11.535	13.130
2	9.188	3.983	6.507	8.181	3.342*	11.544	13.130
3	9.188	1.888*	5.784	8.144	2.713	.002	13.130
4	9.186	9.534	6.314	8.313	6.087*	11.538	12.825
5	9.196	8.943	7.194	8.354	5.708*	11.515	10.205
6	9.206	1.902*	5.791	8.203	2.732	11.618	13.130
7	9.186	11.051	6.274	8.874	8.831*	11.565	12.979
8	9.284	10.974	10.869	8.974	8.617*	11.479	12.991
9	9.382	10.732	11.819	9.072	8.384*	11.420	13.007
10	9.386	2.049*	5.859	8.584	2.936	11.833	12.971
11	9.186	11.265	6.273*	9.774	10.269	13.105	13.131
12	10.164	11.263	11.503	9.914*	10.127	11.442	13.131
13	10.653	11.258	11.410	9.993*	10.047	11.258	13.131
14	11.186	3.348*	6.300	10.146	4.972	4.131	2.135
15	9.186	11.295	6.273*	10.693	11.083	13.131	13.131
16	18.964	11.296	11.385	10.812*	11.012	11.863	13.131
17	23.853	11.297	11.482	10.878*	10.964	765.1169	13.131
18	29.186	8.078	6.417*	13.375	24.438	21.648	13.131
19	9.186	11.300	6.274*	11.201	11.277	13.131	13.131
20	106.964	11.301	11.556	11.228*	11.266	15.081	13.131
21	155.853	11.301	11.516	11.242*	11.258	19.712	13.131
22	209.184	10.860	6.295*	11.811	192.024	89.530	13.131
No. of * scored	0	4	5	6	7	0	0

* = closest to minimum TPSE (attained by $\tilde{\gamma}_{OPT}$); underlined values indicate that the estimator has a smaller TPSE than OLS has with a probability greater than .9.

FIGURE 4.4.4


ESTIMATED TOTAL PSE
RELATIVE TO THAT OF OLS
RELATIVE M2
2.0
1.8
1.6
1.4
1.2
1.0
.8
.6
.4
.2
.0
SMRII
MG
HKB
NAIVE
LW
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
r²=1
r²=10
r²=100
r²=10³
r²=10⁴
r²=10⁵
10-FACTOR MODEL
Θ = OPTIMAL

§4.5 Real Data Analysis

To compare the performance of ridge estimators with the conventional Preliminary Test Estimators (PTE) and OLS, Christ's (1966) investment equations were re-estimated. The data are reproduced in Table 4.5.1 so that cross examination and verification by other investigators are easily allowed. Also we believe that the way Christ formulated and estimated his investment equations is typical of the "Post-Data-Model-Evaluation" strategy as practised by most econometricians. Christ proposed to explain gross investment i by the following linear function (a priori sign of the coefficient are placed on top of the corresponding explanatory variable)

$$i = f(\overset{+}{i_{-1}}, \overset{+}{x_b}, \overset{+}{p}, \overset{+}{a_{-1}}, \overset{-}{a_{-2}}, \overset{-}{k_{-1}}, \overset{+}{d}) \qquad (4.38)$$

where

i = real gross private domestic *investment*

i_{-1} = i lagged one period

x_b = real gross *business product*

p = real *property income*

a_{-1} = *stock market price* index on December 31 of the preceding year

a_{-2} = a_{-1} lagged one period

k_{-1} = real net private domestic *capital stock* on December 31 of the preceding year

d = real capital consumption at replacement cost (*depreciation*).

Note that the model now involves a lagged dependent variable and the sampling theory of adaptive ridge estimators is complicated even with serially independent errors. Christ experimented with 24 different forms of the investment equation by dropping one or more variable from the list

Table 4.5.1

Data Used for Estimation and Testing of Investment Equation

(Source: Christ (1966), p.590-591)

Time t	i_{-1}	x_b	p	a_{-1}	a_{-2}	k_{-1}	d	Gross Investment i
1	27.4	171.5	36.2	40.26	30.85	0.0	22.1	35.0
2	35.0	153.7	25.3	37.28	40.26	12.9	21.0	23.6
3	23.6	142.0	12.8	28.00	37.28	15.5	20.1	15.0
4	15.0	119.3	-0.2	16.91	28.00	10.4	19.0	3.9
5	3.9	115.1	-0.9	15.19	16.91	-4.7	18.3	4.0
6	4.0	125.2	5.8	22.56	15.19	-19.0	18.3	7.4
7	7.4	138.7	14.3	19.74	22.56	-29.9	18.2	16.1
8	16.1	156.6	17.0	27.51	19.74	-32.0	19.6	21.0
9	21.0	167.8	21.4	35.77	27.51	-30.6	20.0	27.0
10	27.0	158.1	17.5	22.26	35.77	-23.6	19.8	15.5
11	15.5	172.1	20.4	26.06	32.26	-27.9	19.8	21.6
12	21.6	188.1	26.4	25.72	26.06	-26.1	20.4	29.0
13	29.0	216.1	32.1	21.53	25.72	-17.5	22.8	36.7
14	17.0	252.6	39.8	25.49	19.70	-46.3	22.7	42.4
15	42.4	259.5	36.7	20.28	25.49	-26.6	26.0	41.5
16	41.5	270.3	45.5	18.11	20.28	-11.1	29.0	49.8
17	49.8	268.8	41.3	17.16	18.11	9.7	30.8	38.5
18	38.5	293.3	42.7	18.75	17.16	17.4	32.2	55.9
19	55.9	311.0	44.2	22.07	18.75	41.1	36.2	57.7
20	57.7	320.3	42.1	24.33	22.07	62.6	36.8	50.4
21	50.4	336.2	39.0	26.54	24.33	76.2	38.3	50.6
22	50.6	330.8	38.5	25.08	26.54	88.5	40.6	48.9
23	48.9	360.5	42.7	34.97	25.08	96.8	42.9	62.5
24	62.5	368.2	41.8	44.83	34.97	116.4	45.6	61.7
25	61.7	375.4	42.1	44.40	44.83	132.5	48.0	58.1
26	58.1	367.6	42.8	37.20	44.40	142.6	47.3	48.3
27	48.3	394.2	45.2	48.48	37.20	143.6	49.1	60.9
28	60.9	405.2	43.0	52.45	48.28	155.4	49.6	60.2
29	60.2	412.0	44.7	49.74	52.45	166.0	49.3	57.5
30	57.5	437.6	49.2	62.00	49.74	178.1	53.1	65.2

of explanatory variables according to a series of preliminary t-tests. His dissatisfaction with *all* the equations are borne out by his remarks on p.601,

> "... There are many a priori wrong signs among the estimates ..., and even where the signs are not wrong some coefficients are not very "significant" ... none of the twenty-four investment is satisfactory in every respect. ... some are more unsatisfactory than others and can be dismissed quickly."

He categorically stated his strategy in arriving at a "satisfactory" equations as follows:

(1) eliminate all equations that have wrong signs

(2) eliminate those equations that have insignificant coefficients

(3) retain those equations that contains "helpful" variables even if they are not significant.

With all this informal use of a priori information, he ended up with two most "reasonable" investment equations, namely

$$\underset{(i15)}{} \quad i = \underset{(3.03)}{.107x_b} + \underset{(4.25)}{.613p} + \underset{(2.27)}{.266a_{-1}} - \underset{(-2.11)}{.275a_{-2}} - \underset{(-.28)}{.0104k_{-1}} - 7.71844$$

$$R^2 = .97$$

and

$$\underset{(i17)}{} \quad i = \underset{(2.50)}{.283i_{-1}} + \underset{(7.21)}{.193x_b} + \underset{(3.55)}{.511a_{-1}} - \underset{(-2.00)}{.338a_{-2}} - \underset{(-4.27)}{.127k_{-1}} - 21.9042$$

$$R^2 = .96$$

where t-values are in parentheses. 27 observations were used over the periods 1929-1941 and 1946-1959.

A casual examination of the correlation matrix of explanatory variables convinces us that there are strong positive linear relationships

between i_{-1} and x_b (.90); i_{-1} and p (.84); x_b and p (.89); k_{-1} and d (.94). The seven e-roots of X'X matrix (after standardisation) are

$$\Lambda = \text{diag}\{4.895, 1.313, .401, .295, .071, .022, .003\}$$

giving a *conditioning ratio*

$$\frac{\lambda_1}{\lambda_p} = \frac{4.895}{.003} \doteq 1632 .$$

It is possible that collinearity will cause OLS estimates to be unreliable. To confirm this and to justify a "declining delta" shrinkage scheme as opposed to a uniform shrinkage scheme, we follow Obenchain (1975) and test the hypotheses

$$H_o: \gamma_1^2\lambda_1 = \gamma_2^2\lambda_2 = \ldots = \gamma_7^2\lambda_7 . \qquad (4.39)$$

It is noted that the rank correlation between spread $\sqrt{\lambda_i}$ and the absolute principal correlates $|r_{yi}|$ (i = 1 ... 7) is greater than .9, indicating that a "declining delta" ridge estimator would be beneficial. Comparing the likelihood ratio statistic of 85.4 with the asymptotic chi-square critical value on 6 degrees of freedom, we can confidently reject H_o in favour of a "declining delta" scheme (i.e. a version of ordinary ridge estimator is preferred to a Stein-type estimator.) Nevertheless, we have calculated the estimates for OLS, Stein positive part rule, the truncated principal components (PC4) (i.e. only 4 of the 7 principal components account for more than 95% of variation) as well as three variants of the five adaptive ridge estimators studied in the last section. They are $\tilde{\beta}_{NAIVE}$, $\tilde{\beta}_{HKB}$ and $\tilde{\beta}_{LW}$ chosen on grounds of simplicity and reliably good performance as compared to the remaining two. The estimates are tabulated in Table 4.5.2 together with the two "reasonable" Christ estimates. Note that the t-statistics are not

Table 4.5.2

Estimates Using Biased Estimators

Variable \ Competitors	i15	i17	OLS	STEIN+	PC4	NAIVE	HKB	LW
i_{-1}	*	.283	.0968	.0921	.1713	.1153	.0929	.0883
x_b	.107	.193	.1304	.1240	.0528	.0861	.0797	.0873
p	.613	*	.5295	.5037	.6092	.5916	.5760	.5695
a_{-1}	.266	.511	.3263	.3104	.3655	.2977	.2939	.3011
a_{-2}	-.275	-.338	-.3341	-.3179	-.4367	-.3376	-.3218	-.3201
k_{-1}	-.0104	-.127	.0022	.0021	-.0176	-.0081	-.0250	-.0255
d	*	*	-.3325	-.3163	.2278	.0070	.1952	.1458
CONSTANT	-7.6184	-21.9042	-4.8182	-2.8108	-5.1396	-5.3473	-7.9112	-8.2145
							k=.0333	k=.0097

(* means the variable has been excluded)

reported as they are not relevant as far as decision theory is concerned.

It is often said that the ridge trace is an essential and informative part of the exercise in ridge analysis. We have therefore carried out a ridge trace along the m-scale proposed by Vinod (1974). Some justifications for the use of the m-scale are as follows:

m is defined as a measure of "multicollinearity allowance"

$$m = p - \delta_1 \ldots - \delta_p \tag{4.40}$$

in a sense that if the δ_i are all zero (total shrinkage) we have $m = p$, while if δ_i are all unity (no modification of OLS) we have $m = 0$. That is, the larger the m, the more serious is the problem of multicollinearity in the data. As $0 < \delta_i < 1$ $(i = 1 \ldots p)$, we necessarily have $0 \leq m \leq p$ so that the δ_i's can be monitored according to some optimality criteria within this range of values of m. Usually, a search in steps of 0.5 is sufficient. This m-scale ridge trace has the following advantages over the well known k-scale ridge trace of Hoerl and Kennard (1970).

(i) Generality and Comparability

The k-scale ridge trace restricts to the 1-parameter ordinary ridge family, while the m-scale allows several different ridge families to be plotted in such a way that the traces are "comparable".

(ii) Measure of Rank Deficiency

As the Principal Components Estimator with "assigned rank" r involves setting the first r deltas $\delta_1 = \ldots = \delta_r = 1$ and the last $p - r$ deltas to zero, so that

$$m = p - r$$

measured the deficiency in rank of the X matrix.

(iii) Linear Stability

If the estimated coefficients, $\tilde{\beta}$, stabilizes, then the right hand tail of the trace will be approximate straight lines all of which intersect at $m = p$ ($\tilde{\beta}_i = 0$, $i = 1 \ldots p$). This is not the case in the k-scale.

(iv) Bayesian Interpretation

Let S = sample variance $= \sigma^2 G\Lambda^{-1}G'$

R = prior variance $= \sigma^2 GK^{-1}G'$

where

$\hat{\gamma} \sim N(\gamma, \Lambda^{-1})$

$\gamma \sim N(0, K^{-1})$

$\beta = G\gamma$.

Thus, the posterior variance is well known to be

$$(S^{-1} + R^{-1})^{-1} = G(\Lambda^{-1} + K^{-1})^{-1}G' .$$

Theil's (1963) relative share is

$$f(S,R) = \frac{1}{p} \text{ tr } S^{-1}(S^{-1} + R^{-1})^{-1}$$

$$= \frac{1}{p} \text{ tr } G\Lambda(\Lambda + K)^{-1}G' .$$

If $\delta_i = \dfrac{\lambda_i}{\lambda_i + k_i}$, $\Lambda = \text{diag}\{\lambda_1 \ldots \lambda_p\}$, $K = \text{diag}\{k_1 \ldots k_p\}$ and $\Delta = \text{diag}\{\delta_1 \ldots \delta_p\} = \Lambda(\Lambda + K)^{-1}$, then we have from (4.40)

$$f(S,R) = \frac{1}{p} \text{ tr } G\Delta G' = \frac{p-m}{p} . \qquad (4.41)$$

For instance, $m = 3$ and $p = 10$, $f(S,R) = \dfrac{7}{10}$ implying that 70% of the posterior precision is due to sample information.

The numerical ridge trace is given in Table 4.5.3 estimating the model with all possible explanatory variables included. This reports selected points on the Curve Dé-colletage of Dickey (1974) and reveals the sensitivity of the location of the posterior distribution to a slight change of the prior variances within the class of spherical normal priors. Perhaps by so doing, it has come closer to meeting the demand for good reporting of Bayesian studies. (See Roberts (1974).) It is clear that the column under $m = k = 0$ are OLS estimates. The pattern of ridge estimates shows how the wrongly signed coefficients are 'corrected' as one goes from the left to the right of the trace, but up to a point the bias introduced was too severe that the estimates do not make sense any more. Certain criteria must be called to pick up the "right" ridge estimate from the trace. We have already mentioned the Mallow's C_p, the SSCBC and the $-2 \ell n L$ criteria. Minima of these quantities along the trace were underlined indicating the corresponding columns of $\tilde{\beta}_{CP}$, $\tilde{\beta}_{SSCBC}$ and $\tilde{\beta}^{+}_{OPT}$ respectively. The signs of these estimates are of the expected 'correct' ones and the magnitudes of the coefficient appear to be more 'reasonable' than those of OLS. Consider the short run effect of output (coefficient of x_b), given by $\hat{\beta}$ and $\tilde{\beta}_{SSCBC}$, are .1304 and .0597 respectively, while the long run effect are $\frac{.1304}{1-.0968} = .1444$ and $\frac{.0597}{1-.1364} = .0691$ which must then result in two drastically different policy proposals by economists using the two different sets of estimates. The graphical ridge trace is displayed at Figure 4.5.1 and the minimum SSCBC does indeed locate the point where the ridge estimates begin to stabilise. The trace reveals dramatic change in sign and magnitude for the coefficient of d and a slight change of the coefficient of k_{-1}. This may be due to the strong correlation between the two variables.

We have in hand all the basic information from a ridge analysis of the investment model. To answer the hard question of which particular

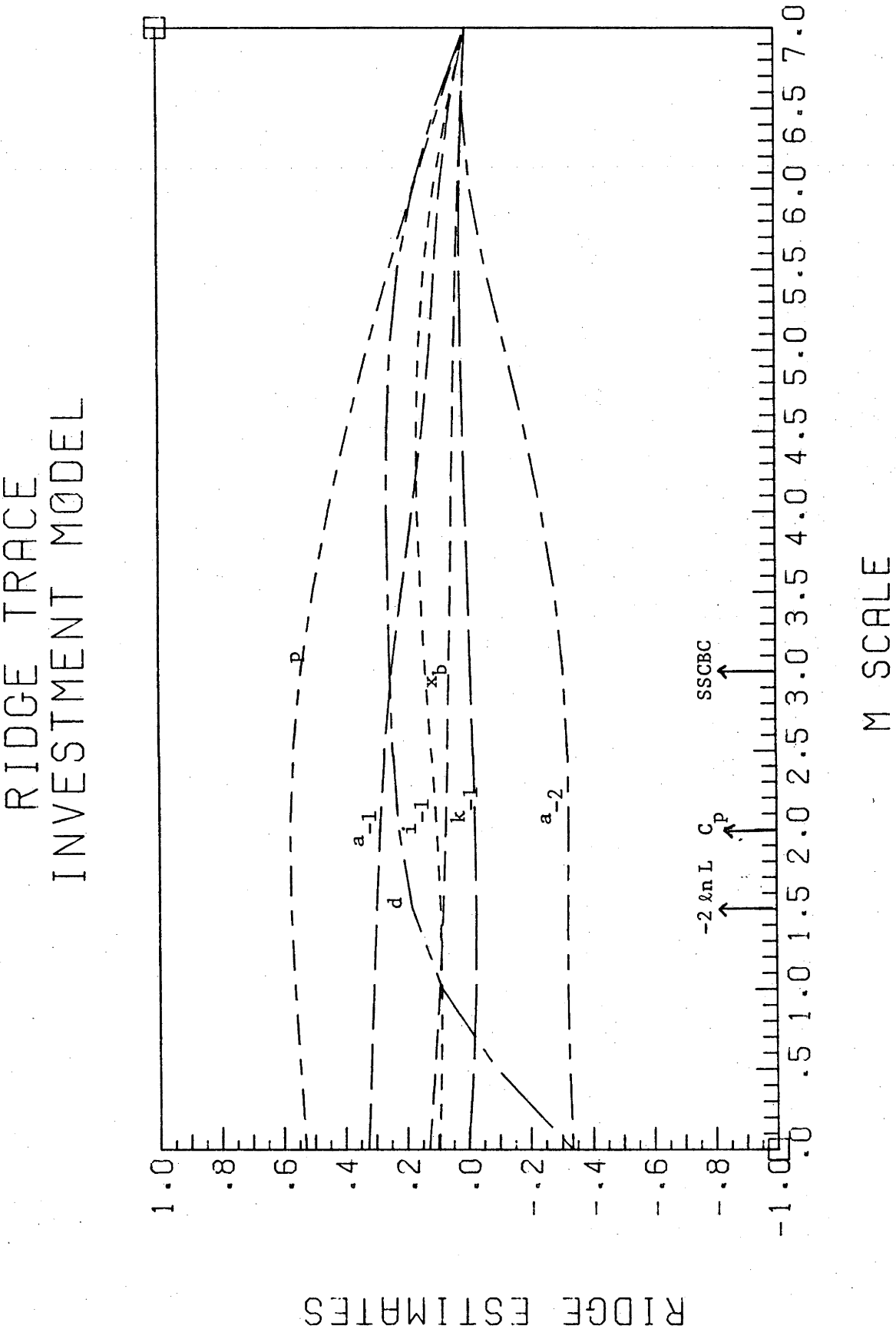
TABLE 4.5.3

Ridge Trace for Estimates of Investment Equation
(Data Source: Christ (1966) Ch.11)

Rank deficiency measure m		OLS	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0	6.6	7.0
Ridge parameter k		0	.0019	.0060	.0147	.0311	.0602	.1099	.1930	.3318	.5701	1.0034	1.8669	3.8762	10.5633	∞
EXPLANATORY VARIABLES AND A PRIORI SIGNS	i_{-1}	.0968 (.79)	.0900	.0872	.0911	.1023	.1187	.1364	.1513	.1607	.1626	.1550	.1360	.1039	.0580	0
	x_b	.1304 (1.76)	.1107	.0942	.0819	.0726	.0654	.0597	.0547	.0502	.0454	.0398	.0328	.0239	.0129	0
	p	.5295 (2.90)	.5461	.5624	.5745	.5767	.5650	.5388	.4992	.4483	.3882	.3207	.2473	.1688	.0861	0
	a_{-1}	.3263 (2.29)	.3163	.3063	.2962	.2845	.2682	.2440	.2121	.1769	.1445	.1198	.1024	.0839	.0523	0
	a_{-2}	-.3341 (-2.17)	-.3244	-.3200	-.3212	-.3239	-.3218	-.3091	-.2825	-.2417	-.1886	-.1266	-.0637	-.0132	.0102	0
	k_{-1}	.0022 (.03)	-.0143	-.0238	-.0255	-.0221	-.0167	-.0101	-.0028	.0049	.0123	.0183	.0213	.0198	.0126	0
	d	-.3325 (-.36)	-.0887	.0874	.1830	.2249	.2424	.2508	.2563	.2595	.2578	.2465	.2196	.1710	.0979	0
	Constant	-4.8182 (-.40)	-6.9781	-8.1168	-8.0479	-7.1739	-5.9010	-4.4016	-2.7128	-.8066	1.4262	4.2703	8.3285	14.5507	23.8453	36.3878
RIDGE SELECTION CRITERIA	$1-R^2$=ESS(·)	.026	.026	.026	.026	.027	.028	.031	.038	.049	.070	.107	.176	.306	.555	1
	Mallow's C_p	8	7.0752	6.2525	5.5076	4.9726	4.9883	6.2124	9.7576	17.3420	31.7623	58.2880	107.8417	203.1079	385.6628	712.1871
	SSCBC(·)	4.0663	3.2504	2.7593	2.3932	1.8834	1.4007	1.2605	1.4165	1.5867	1.6999	1.9735	2.7809	4.4147	6.7982	9.4648
	-2 ln L(·)	∞	15.4415	-82.3491	-93.3754	-87.1313	-77.0777	-66.8143	-57.4497	-49.2280	-42.0500	-35.7340	-30.1221	-25.0677	-20.2724	-12.6326

Minimum value of selection criteria are underlined.
t-ratios of OLS estimates are in parenthesis.

FIGURE 4.5.1


RIDGE TRACE
INVESTMENT MODEL
RIDGE ESTIMATES
M SCALE
1.0
.8
.6
.4
.2
.0
-.2
-.4
-.6
-.8
-1.0
.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
p
a_{-1}
d
i_{-1}
x_b
k_{-1}
a_{-2}
SSCBC
$-2 \ln L$
c_p

set of point estimates to choose, we must resort to prior information or experience. Of course, the test of the pudding is in the eating. Since we have left the last three observations for model testing we can compare the various estimators in term of their ability to predict gross investment. The forecasts and their errors are given in Table 4.5.4 and they are graphed in Figure 4.5.2. Note that we have omitted the "naive" ridge estimator from comparison, because we know that it is not very different from either $\tilde{\beta}_{HKB}$ or $\tilde{\beta}_{LW}$.

From Table 4.5.4, it is interesting to note that the different estimators can be grouped according to the size of their total prediction squared error. $\tilde{\beta}_{PC4}$ and $\tilde{\beta}_{SSCBC}$ have a TPSE < 10 and are seen to outperform the remaining estimators by a big margin. Next is the group including $\tilde{\beta}_{CP}$, $\tilde{\beta}_{STEIN+}$, $\tilde{\beta}_{HKB}$, $\tilde{\beta}^{+}_{OPT}$ (or $\tilde{\beta}_{-2\ \ell n\ L}$) and $\tilde{\beta}_{LW}$, which have TPSE between 10 and 20. Although they lose out to the first group, these rules beat OLS and the PTE's of Christ handsomely. One can judge for oneself the wisdom of using OLS or PTE in the analysis of non-experimentally generated data.

§4.6 Conclusion

The major contribution of this chapter is to compare the potential and realised improvement of alternative estimators on OLS. There have been numerous simulations comparing the performance of biased estimators in the literature, however, these have not compared all the biased alternatives considered in this chapter in one unified study. To the best of our knowledge, we think that our comparison of Stein positive part rule, principal components estimator and various versions of adaptive ridge estimator using the carefully planned designs of Baldwin and Hoerl (1978) is the first of its kind in this area. The result indicates when and where a particular biased estimator should be expected to perform well

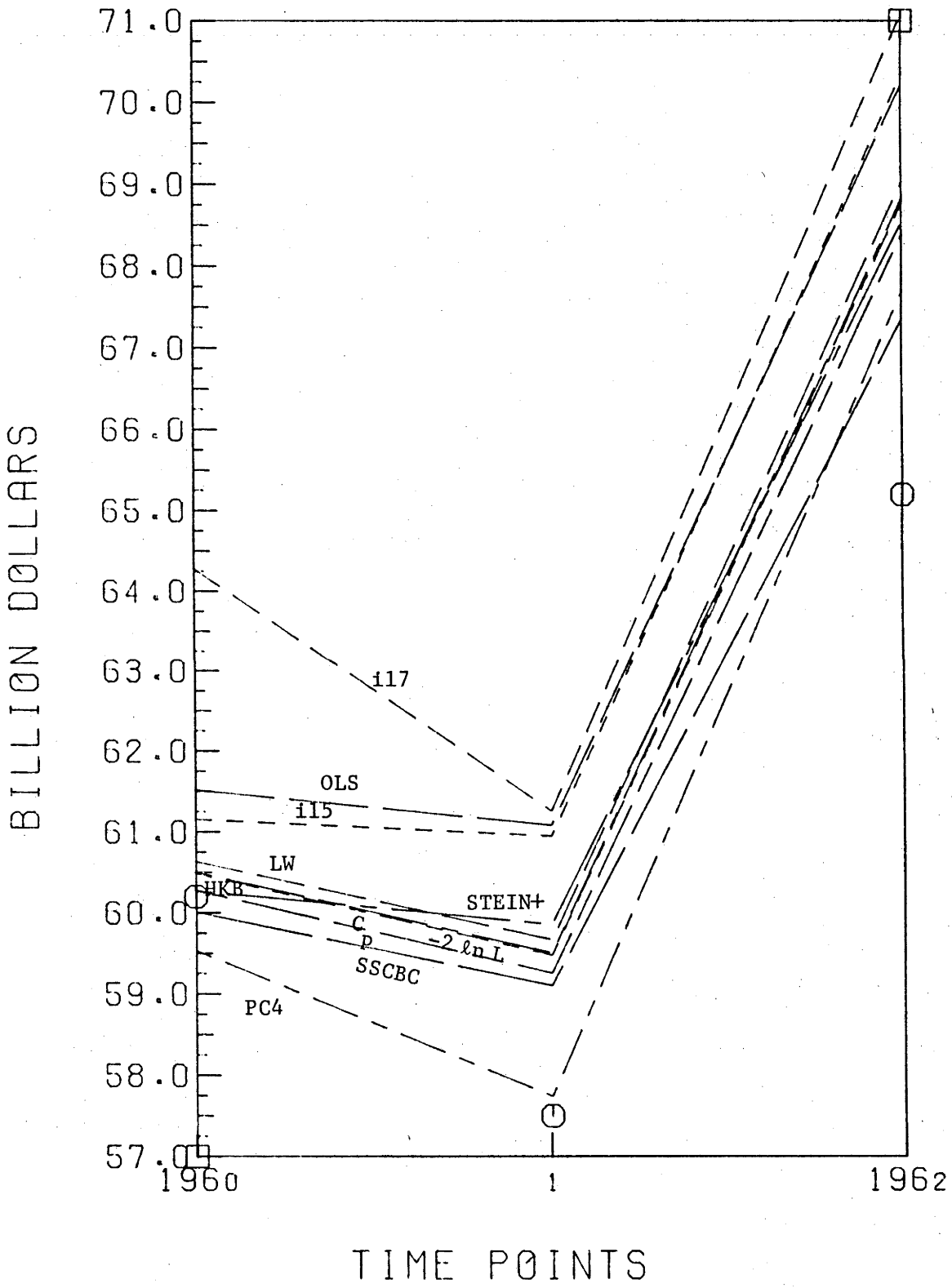
Table 4.5.4

Post-Sample Forecasting Comparison

Estimators \ Time Points	1960	1961	1962	Total predictor square error
Observed Investment	60.2	57.5	65.2	0
PC4	59.533	57.742	67.689	6.799
SSCBC	60.009	59.103	67.347	7.215
C_p	60.277	59.256	68.386	13.240
STEIN+	60.272	59.858	68.540	16.721
HKB	60.484	59.479	68.775	16.773
$-2 \ell n L$	60.502	59.504	68.829	17.277
LW	60.633	59.668	69.032	19.567
i15	61.156	60.947	70.326	39.010
OLS	61.518	61.082	70.211	39.676
i17	64.282	61.255	71.076	65.294

FIGURE 4.5.2

FORECASTING OF INVESTMENT YEARS 1960 1961 AND 1962



⏲ = OBSERVED INVESTMENT

and to exploit most of the potential gain promised by the 'existence' theorems of the optimal ridge estimator. The negative results obtained by Newhouse and Oman (1971) about the performance of ridge estimator can be assessed on the basis of our experimental evidence. The fact that Newhouse and Oman based their experiment on models of size $p = 2$ only did not appropriately reflect the regression problems tackled by econometricians in general. Also, they kept the error variance σ^2 constant at $\frac{1}{9}$ and the true coefficient vector length at $|\beta|^2 = 1$ effectively restricting the signal-noise ratio to be as high as 9 throughout the experiment. Comparing this with the scaled TMSE $\left(\sum_{i=1}^{p} \frac{1}{\lambda_i} \right)$ of OLS estimators in their designs which were equal to 11.734 (when collinearity $r = .91$) and 137.488 (when collinearity $r = .99$) respectively, it is not surprising that ridge estimator did not do too well in the first case but did considerably better than OLS when the true vectors are oriented to the leading e-vectors in the latter case. This confirms Baldwin and Hoerl's result that one should only apply ridge estimator if one has reason to believe that the signal-noise ratio is much smaller than the scaled TMSE of OLS in a particular specification. Our Monte Carlo results also support this observation and in the low signal-noise ratio region, the orientation of the true vector will not be an overly important factor.

Relatively few applications have been made of the biased estimators to real economic analysis. Aigner and Judge (1975) applied Stein positive part rules to a number of economic models and found that the gain from using these estimators over the traditional ones of PTE or OLS is trivial. They based their comparison, as we have done in our investment example, on out-of-sample period forecast. One possible reason for the apparent good performance of Stein positive rule in our case but not-so-good

performance in theirs may be due to the collinearity structure in their data. Unfortunately, they did not provide the data used in their analysis so we are unable to verify their conclusions as well as to find out if a version of adaptive ridge estimator would produce significant gain such as those achieved in our investment example. Vinod (1974) and Vinod (1976) reported successful applications of ridge estimators in estimating production functions. Maddala (1977) applied the iterative version of $\tilde{\beta}_{HKB}$ to re-estimate the distributed lag model on Almon data and found that the ridge estimated coefficients were all almost equal. He found that the Shiller method gave more "reasonable" estimates. However, his refutation of ridge estimator is based on a purely subjective argument. If we are to compare the various estimators on the basis of total prediction errors, we would expect a form of adaptive ridge estimator (without iteration) to do well. Guilkey and Murphy (1975) applied their Directed Ridge Regression technique to Christ's investment data and their justification is also subjective. Since their estimates were computed from a restricted model, it is expected that their forecasting ability would be inferior to our ridge estimates based on the "full" model. Brown and Beattie (1975) observed that for economic models with positive intercorrelation among explanatory variables, the mean square errors of ridge estimate will be relatively small if the true coefficients all have the same sign and approximately the same magnitude. This reflects the Bayesian view of ridge estimator as sort of posterior mean with respect to an exchangeable normal prior distribution. They applied ridge estimation technique to estimating the marginal value productivity of water in irrigated agriculture of Western United States. Comparison was made with OLS estimates of a 6-variable Cobb-Douglas production function. Their ridge estimates were identified by some subjective rules along a k-scale ridge trace and "controlled" by a Wallace-type weak mean square error test to avoid an over-biased ridge estimate

being selected. In their results, the success of ridge estimate were also judged by a priori information. It would be more convincing if they had also carried out post sample period forecast.

In summary, a suitable choice of adaptive ridge estimator is likely to produce significant gains over traditional estimation techniques. Our results of Truncated Principal Components Estimator performing comparably well relative to some 'good' adaptive ridge estimator is not surprising if we consider the similarity of the shrinkage deltas for $\tilde{\beta}_{PC4}$ and for $\tilde{\beta}_{SSCBC}$. They are given respectively by

$$\Delta_{PC4} = \{1, 1, 1, 1, 0, 0, 0\}$$

$$\Delta_{SSCBC} = \{.978, .923, .785, .728, .393, .166, .026\} .$$

Our Monte Carlo results indicate that ridge estimator should perform well under the criteria of total mean squared errors and/or total prediction squared errors if the signal-noise ratio are less than $\sum_{i=1}^{p} \frac{1}{\lambda_i}$. Of course, if a priori information is not compatible with the exchangeability assumption, some modification of the ordinary ridge estimators, such as a Shiller-type approach would give more "plausible" point estimates. Nevertheless, this does not prevent a straight ridge estimator or a truncated principal component estimators from performing well under the loss functions we commonly use.

CHAPTER 5

"IMPROVED" ESTIMATORS IN ECONOMETRICS I: DISTRIBUTED LAG MODELS

§5.1 Introduction

It is generally acknowledged that economic phenomena are dynamic. Owing to technical, institutional and psychological rigidities, some lapses of time are required for economic variables to adjust fully to an external shock in the system. Economists distinguish between short run and long run reactions of, say, consumer expenditure to an increase in disposable income or that of investment expenditure to a change in sales volume. They also want to know when a certain economic measure takes effect and when this effect is fully worked out. Irving Fisher (1925) was the first econometrician to introduce and discuss the concept of distributed lags. However, inadequacy of statistical methods and data in those early days thwarted the efforts of Alt (1942) and Tinbergen (1949) to popularise the idea. Since the publication of Koyck's (1954) systematic study of estimation problems relating to this model, there has been considerable research efforts devoted to developing related estimation techniques and applying them in a variety of empirical contexts. The contributions are now of such a scale that it is very difficult to be exhaustive on the subject in a single survey. Dhrymes' (1971) treatise gives an extensive account on the statistical foundation of various estimators. A more elementary survey is given by Wallis (1969). More recent surveys include the work of Sims (1974), Richard (1977) and Hendry and Pagan (1979). This chapter will deal with two things:

(1) to study the problems arising from applying least square procedure to estimate the general distributed lag model

$$y_t = \sum_{i=o}^{\infty} \beta_i x_{t-i} + \varepsilon_t \qquad (5.1)$$

and

(2) to examine possible alternative estimators both in the time and frequency domains.

The Model (5.1) in this form is unidentifiable because there are too many unknown parameters to be estimated. Various approximations to (5.1) have to be made so that estimation of the model based on a set of finite number of observations on y and x is feasible.

Writing the Model (5.1) as

$$y_t = \beta(L)x_t + \varepsilon_t \qquad (5.2)$$

where $\beta(L)$ is a polynomial in the lag operator L of infinite order, the most common form of approximation to (5.2) is by means of Jorgenson's (1963) Rational Distributed Lag (RDL) structure where the general infinite lag function $\beta(L)$ is to be replaced by a ratio of two finite polynomials in L, viz.

$$\beta(L) \simeq \frac{\gamma(L)}{\alpha(L)} = \frac{\gamma_o + \gamma_1 L + \ldots + \gamma_p L^p}{\alpha_o + \alpha_1 L + \ldots + \alpha_q L^q} \qquad (5.3)$$

where $\gamma(L)$ and $\alpha(L)$ are relatively prime. Any lag function can be approximated by a rational lag function as closely as we like by making p and/or q sufficiently large. Substituting (5.3) into (5.2) and rearranging, we have

$$\alpha(L)y_t = \gamma(L)x_t + \alpha(L)\varepsilon_t \ . \qquad (5.4)$$

This equation, which is a special case of the general ARMAX model [Nicholls, Pagan and Terrell (1975)], contains q lagged values of dependent variable and p lagged values of the explanatory variables.

The error term is seen to be a q-order moving average of the original disturbance term. If ε_t is random, $\alpha(L)\varepsilon_t$ will be serially correlated and OLS will be inconsistent. Thus, in general, alternative estimators may be required. Estimation problems aside, we note that quantities of economic interest such as the total multiplier (or the long run equilibrium effect on y of a unit change in x) and the mean lag are given by (5.5) and (5.6) respectively.

$$\text{Total Multiplier} = \beta_o + \beta_1 + \beta_2 + \ldots = \beta(1) \simeq \frac{\gamma(1)}{\alpha(1)} \tag{5.5}$$

$$\text{Mean Lag} = \frac{\sum_{i=o}^{\infty} i\beta_i}{\sum_{i=o}^{\infty} \beta_i} = \frac{\dot{\beta}(1)}{\beta(1)} \tag{5.6}$$

where $\dot{\beta}(1) = \left.\frac{d\beta(L)}{dL}\right|_{L=1}$ and $\beta_i \geq 0$ all i. Sims (1974) pointed out that the concept of mean lag could be misleading if the distribution of the weights β_i $(i = 1 \ldots \infty)$ is not symmetric and alternative measure such as the median is usually more suitable. Hence, care must be taken to interpret the quantity at (5.6).

Recently, Nerlove (1972) and other econometricians have complained about the lack of economic justifications for the various approximations to (5.1). It appears that the most convincing argument for the model (5.1) comes from the rational expectation literature. Muth (1961) argued that in forming their expectations, economic agents take account of the inter-relationships among variables rather than simply extrapolating from one variable's past history. Consider a "classical" static model

$$By_t + \Gamma_1 y_{1t}^* + \Gamma_2 x_t = u_t \ . \tag{5.7}$$

The reduced form of the structure (5.7) is

$$y_t = -B^{-1}\Gamma_1 y_{1t}^* - B^{-1}\Gamma_2 x_t + B^{-1}u_t \ . \tag{5.8}$$

Partition (5.8) and rewrite as

$$y_t = \begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} \Pi_{11} & \Pi_{12} \\ \Pi_{21} & \Pi_{22} \end{pmatrix} \begin{pmatrix} y^*_{1t} \\ x_t \end{pmatrix} + \begin{pmatrix} v_{1t} \\ v_{2t} \end{pmatrix} . \qquad (5.9)$$

Taking conditional expectations of y_{1t}, given an information set Ω_{t-1}, say, we have

$$y^*_{1t} = \mathcal{E}(y_{1t}|\Omega_{t-1}) = \Pi_{11}y^*_{1t} + \Pi_{12}\mathcal{E}(x_t|\Omega_{t-1})$$

i.e.

$$y^*_{1t} = (1-\Pi_{11})^{-1}\Pi_{12}\mathcal{E}(x_t|\Omega_{t-1}) . \qquad (5.10)$$

Assuming that the list of exogenous variables in the model is correct and complete, $\mathcal{E}(x_t|\Omega_{t-1})$ can be optimally predicted by its past history $x_{t-1}, x_{t-2}, \ldots$. Thus, the "observable" reduced form equations of (5.8) exhibit each endogenous variables as a distributed lag function of the exogenous variables. Of course, unlike the usual data-based distributed lag model, this distributed lag function assumes a special form whose weights must satisfy suitable restrictions. [See Wallis (1977)]. Many other economic justifications, somewhat ad hoc, are summarised in Hendry and Pagan (1979).

In the ensuing sections, we assume that model (5.1) has been theoretically justified and shall only be concerned with the empirical identification of the lag weights of the model. Various alternative estimators are examined and the techniques of some good estimation procedures are extended.

§5.2 Estimation in Time Domain

The rational distributed lag model with $\gamma(L)$ of order p and $\alpha(L)$ of order q can be denoted by RDL(p,q). It is easy to see that given the coefficients of $\gamma(L)$ and $\alpha(L)$, the weights of $\beta(L)$ are restricted as

$$\beta(L)\alpha(L) = \gamma(L) \ . \tag{5.11}$$

Expanding and equating coefficients, we have the 'convolution' identities:

$$\begin{aligned}
&\text{constant term} && : \gamma_o = \beta_o\alpha_o \\
&\text{coefficient of } L && : \gamma_1 = \beta_o\alpha_1 + \beta_1\alpha_o \\
&\text{coefficient of } L^2 && : \gamma_2 = \beta_o\alpha_2 + \beta_1\alpha_1 + \beta_2\alpha_o \\
&\qquad\vdots \\
&\text{coefficient of } L^q && : \gamma_q = \beta_o\alpha_q + \beta_1\alpha_{q-1} + \ldots + \beta_q\alpha_o \\
&\text{coefficient of } L^{q+1} && : \gamma_{q+1} = \beta_1\alpha_q + \ldots + \beta_{q+1}\alpha_o \\
&\text{coefficient of } L^{q+2} && \gamma_{q+2} = \beta_2\alpha_q + \ldots + \beta_{q+2}\alpha_o \\
&\qquad\vdots \\
&\text{coefficient of } L^p && : \gamma_p = \beta_{p-q}\alpha_q + \ldots + \beta_p\alpha_o \\
&\text{the rest} && \begin{cases} 0 = \beta_{p-q+1}\alpha_q + \ldots + \beta_{p+1}\alpha_o \\ \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ 0 = \alpha(L)\beta_t \qquad\qquad t \geq p+1 \end{cases}
\end{aligned}$$

Since $\gamma_o \ldots \gamma_p$ are not restricted in any way, the lag weights $\beta_o \ldots \beta_p$ are not restricted. However, $\beta_{p+1}, \beta_{p+2}, \ldots$ must follow a q-order autoregressive scheme, i.e.

$$\alpha(L)\beta_t = 0 \qquad t = \overline{p+1,\infty} \ . \tag{5.12}$$

It is a common misconception that whenever $q > 0$, RDL (p,q) must imply an infinite order for $\beta(L)$. Hendry and Pagan (1979) illustrated that using RDL (p,q) for finite distributed lag model is equivalent to imposing end-point restrictions in addition to the implied restrictions.

By far, the most popular RDL (p,q) is for $p = 0$ and $q = 1$, namely, the geometric lag structure. This model has a particularly simple reduced form. From (5.4), we have

$$(\alpha_o+\alpha_1 L)y_t = \beta_o x_t + (\alpha_o+\alpha_1 L)\varepsilon_t \tag{5.13}$$

i.e.

$$y_t = \alpha y_{t-1} + \beta_o x_t + v_t \tag{5.14}$$

where

$$\alpha_o = 1$$

$$\alpha_1 = -\alpha$$

$$v_t = \varepsilon_t - \alpha\varepsilon_{t-1} .$$

Despite its simplicity, problems arising from the estimation of (5.14) are considerable. Because of the simultaneous presence of a lagged dependent variable and a serially correlated error, OLS is well known to be inconsistent in this situation. Attempts to obtain consistent and asymptotically efficient estimate have been made by Hannan (1967) and Wallis (1967). Comparisons of the properties of various estimating procedures have been carried out by Amemiya and Fuller (1967) (analytically) and by Morrison (1970) (simulation). Recently, there is a diversion of attention from the RDL (p,q) models because of the poor forecasting performance when such equations are simulated dynamically together with an autocorrelated error structure. Furthermore, as the RDL model implicitly assumes a geometrically declining lag weights, the distribution of lag weights is then not a symmetrical one and the mean lag obtained for such a distribution can be misleading. An interesting

example given in Hendry and Pagan (1979) is that it is possible that 90% of the adjustment is completed within the first two period in a particular model while it has a mean lag of ten periods. On the other hand, the finite distributed lag, with restrictions on its weights β_i ($i = 0 \ldots n$, say) to lie on a polynomial of low order r, say, and without lagged dependent variables, gives more stable and accurate forecasts. Such a modified model of Almon (1965) was later found to be too artificial and rigid, as one is assuming too much about the lag weights. As a consequence, more recent research on distributed lags aim at relaxing the rigid restrictions imposed by Almon's technique. Consider the Finite Distributed Lag (FDL) model

$$y_t = \beta_o x_t + \beta_1 x_{t-1} + \ldots + \beta_n x_{t-n} + \varepsilon_t \qquad \begin{matrix} t = 1 \ldots T \\ n < \infty \end{matrix} . \tag{5.15}$$

Following Almon (1965), we assume that

$$\beta_i = \alpha_o + \alpha_1 i + \alpha_2 i^2 + \ldots + \alpha_r i^r \qquad \begin{matrix} i = 0 \ldots n \\ n \geq r \geq 0 \end{matrix} . \tag{5.16}$$

Then, equation (5.15) can be written as

$$\begin{aligned} y_t &= \sum_{i=o}^{n} (\alpha_o + \ldots + \alpha_r i^r) x_{t-i} + \varepsilon_t \\ &= \alpha_o x^*_{ot} + \alpha_1 x^*_{1t} + \alpha_2 x^*_{2t} + \ldots + \alpha_r x^*_{rt} + \varepsilon_t \end{aligned} \tag{5.17}$$

where

$x^*_{jt} = \sum_{i=o}^{n} i^j x_{t-i}$ are the Almon variables ($j = 0 \ldots r$).

In matrix notation, the (n+1)-vector of lag weights β are related to the (r+1)-vector of Almon coefficients α as

$$\beta = H\alpha \tag{5.18}$$

where H is the $(n+1)\times(r+1)$ Vandermonde Matrix

$$H = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & \cdots & 1 \\ \vdots & & & \\ 1 & n & & n^r \end{bmatrix} .$$

Since the columns of H are linearly independent, we can define the idempotent matrix

$$M = I_{n+1} - H(H'H)^{-1}H' . \tag{5.19}$$

Thus, it is easy to see that the Almon estimation procedure is the Restricted Least Square (RLS) estimator of β subject to the homogeneous linear restrictions on β

$$M\beta = 0 \tag{5.20}$$

giving

$$\hat{\beta}_{RLS} = \hat{\beta}_{OLS} - (X'X)^{-1}M'[M(X'X)^{-1}M']^{-1}M\hat{\beta}_{OLS} . \tag{5.21}$$

However, one is never sure of the true value of n and r and the implications of misspecification of any one of them are inefficient and/or biased estimator (see Trivedi and Pagan (1979)). Even if the lag length, n, is correctly specified, the lag weights could not be expected to lie exactly on an r-order polynomial. A slightly more flexible approach is to suggest that the a priori information is uncertain.

An efficient way to handle stochastic information is by Bayes method. The Lindley-Smith (1972) hierarchical priors of hyper-parameters are particularly relevant in this situation, which has been reviewed in Section 1.5. To recapitulate, if we specify that the sample of data has a normal distribution

$$y \sim N_T(X\beta, \Omega)$$

while the lag weights has a normal prior

$$\beta \sim N_{n+1}(H\alpha, V)$$

then the posterior mean conditional on V and α is given by

$$\tilde{\beta}_B = (X'\Omega^{-1}X + V^{-1})^{-1}(X'\Omega^{-1}y + V^{-1}H\alpha) \ . \tag{5.22}$$

Of course, α is rarely known, one can then go one step further by assuming a diffuse prior on α, so that the Lindley-Smith Estimator of β under an exchangeable assumption becomes

$$\tilde{\beta}_{EX} = [X'\Omega^{-1}X + V^{-1} - V^{-1}H(H'V^{-1}H)^{-1}H'V^{-1}]^{-1}X'\Omega^{-1}y \ .$$

In the special case where $\Omega = \sigma^2 I_T$ and $V = \sigma_\beta^2 I_{n+1}$, we have

$$\tilde{\beta}_{EX} = [X'X + kM]^{-1}X'y \tag{5.24}$$

where

M is defined in (5.19) and $k = \dfrac{\sigma^2}{\sigma_\beta^2}$

For $H \equiv I$, $\tilde{\beta}_{EX}$ collapses back to the well known Lindley-Smith Estimator at (1.26).

It is instructive to note the similarity between $\tilde{\beta}_{EX}$ and the ridge estimator. The ridge parameter k can be interpreted as the ratio of sample to prior variances $\dfrac{\sigma^2}{\sigma_\beta^2}$. If k equals zero (a diffuse prior for β), $\tilde{\beta}_{EX}$ is simply the OLS estimator, while if $k \to \infty$ (a tight prior for β), $\tilde{\beta}_{EX}$ becomes the restricted least square estimator of Almon. In fact, it is not hard to establish an existence theorem analogous to those for the ridge estimators in Chapter 3.

Theorem 5.2.1

The difference of the risk matrix measure of $\tilde{\beta}_{OLS}$ and that of $\tilde{\beta}_{EX}$ will be positive semidefinite if

$$\text{(i)} \quad \beta'X'X\beta < \sigma^2 \tag{5.25}$$

or

$$\text{(ii)} \quad k < \frac{2\sigma^2}{\beta'\beta} \tag{5.26}$$

or

$$\text{(iii)} \quad \beta'[\frac{2}{k} I + (X'X)^{-1}]^{-1}\beta < \sigma^2 \ . \tag{5.27}$$

Proof.

$$\tilde{\beta}_{EX} = (X'X+kM)^{-1}X'y = Z\hat{\beta}_{OLS}$$

where

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'y \qquad Z = (X'X+kM)^{-1}X'X$$

$$\begin{aligned} MSE(\tilde{\beta}_{EX}) &= \mathcal{E}(\tilde{\beta}_{EX}-\beta)(\tilde{\beta}_{EX}-\beta)' \\ &= \sigma^2 Z(X'X)^{-1}Z' + (Z-I)\beta\beta'(Z-I)' \\ &= (X'X+kM)^{-1}\{\sigma^2X'X + k^2M\beta\beta'M\}(X'X+kM)^{-1}. \end{aligned} \tag{5.28}$$

On the other hand

$$\begin{aligned} MSE(\hat{\beta}_{OLS}) &= \sigma^2(X'X)^{-1} \\ &= \sigma^2(X'X+kM)^{-1}(X'X+kM)(X'X)^{-1}(X'X+kM)(X'X+kM)^{-1} \\ &= \sigma^2(X'X+kM)^{-1}\{X'X+2kM+k^2M(X'X)^{-1}M\}(X'X+kM)^{-1}. \end{aligned} \tag{5.29}$$

Since M is a symmetric idempotent, we have

$$\begin{aligned} &MSE(\hat{\beta}_{OLS}) - MSE(\tilde{\beta}_{EX}) \\ &= k(X'X+kM)^{-1}M\{2\sigma^2I + \sigma^2k(X'X)^{-1} - k\beta\beta'\}M(X'X+kM)^{-1}. \end{aligned} \tag{5.30}$$

For $k > 0$, this expression will be positive semidefinite (p.s.d.) if any of the following holds

(i) $\{\sigma^2(X'X)^{-1} - \beta\beta'\}$ is p.s.d.

i.e. $\beta'X'X\beta < \sigma^2$

or

(ii) $\{2\sigma^2 I - k\beta\beta'\}$ is p.s.d.

i.e. $k < \frac{2\sigma^2}{\beta'\beta}$

both are sufficient conditions for

(iii) $\{\frac{2\sigma^2}{k} I + \sigma^2(X'X)^{-1} - \beta\beta'\}$ is p.s.d.

i.e. $\beta'[\frac{2}{k} I + (X'X)^{-1}]^{-1}\beta < \sigma^2$. □

Remarks: This theorem shows that conditions for MSE-goodness of estimators subject to stochastic Almon Restrictions are the same as those for the straight ridge estimators established in Chapter 3.

Note that the stochastic Almon restrictions $\beta \sim N(H\alpha, V)$ can be improved upon by adjusting H so as to satisfy the assumptions that β_i's lie on a piecewise polynomial. These linear restrictions on β are convenient and computationally easy, but the shapes for the lag distribution could be erratic since this prior does not allow the β_i's to deviate far from some polynomial and is indifferent to the inequalities of the deviations. Shiller (1973) argued that the proper prior should reflect our belief that the lag distribution is smooth. His smoothness prior is based on the idea that if β_i's lie on an r-order polynomial, then the r+1'th successive differencing of the lag weights should be zero. This notion of smoothness can be represented by

$$Rw \sim N_{n-r}(0, \sigma_\beta^2 I) \tag{5.31}$$

where R is a $(n-r)\times(n+1)$ differencing matrix. Then, using the Theil-Goldberger (1961)'s mixed-estimation method, we obtain the Shiller

estimator of β as

$$\tilde{\beta}_S = (X'X + kR'R)^{-1}X'y \tag{5.32}$$

here again $k = \frac{\sigma^2}{\sigma_\beta^2}$ is assumed known.

If $k = 0$, then $\tilde{\beta}_S \equiv \hat{\beta}_{OLS}$, however, if $k \to \infty$, we can use a result in Rao (1973, p.33) to get

$$\tilde{\beta}_S = \hat{\beta}_{OLS} - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R\hat{\beta}_{OLS} \tag{5.33}$$

which can be shown to be equivalent to the non-stochastic Almon estimator (5.21).

In practice, the explanatory variables is an autocorrelated time series so that columns of X are expected to be highly correlated. Although the straight ridge estimator is particularly useful in breaking multicollinearity, the implied prior of that estimator is based on the exchangeable assumption that all lag weights are equal which may contradict our prior knowledge about lag distributions for most of the economic variables. Nevertheless, it can be regarded as a special case of the Shiller estimator with $R = I_{n+1}$. We would therefore expect $\tilde{\beta}_{EX}$ in (5.24) or $\tilde{\beta}_S$ in (5.32) to be a more flexible alternative to the unreliable $\hat{\beta}_{OLS}$ and to the over-restrictive Almon estimator. The remaining problem for using this ridge-type estimator is the selection of a suitable value of k. Of course, one can follow Hoerl and Kennard's (1970) suggestion in selecting the k where the ridge trace begins to stabilise. Alternatively, Maddala (1977) suggested that one can estimate k iteratively by setting

$$k_{(t)} = \frac{\hat{\sigma}^2}{\hat{\sigma}_\beta^2(t)}, \quad \text{where} \quad \hat{\sigma}_\beta^2(t) = \frac{\hat{\beta}^{(t)'}R'R\hat{\beta}^{(t)}}{n+1},$$

$$\hat{\beta}^{(t)} = (X'X + k_{(t-1)}R'R)^{-1}X'y$$

and (t) is the iteration number.

As $\hat{\sigma}^2$ is fixed in each iteration, $k_{(t)}$ will be bigger as $\hat{\sigma}^2_\beta(t)$ gets smaller, it is not hard to see that $\hat{\sigma}^2_\beta(t) < \hat{\sigma}^2_\beta(t-1)$ because $R\hat{\beta}^{(t)}$ is being pulled closer and closer to the null vector. This sequence is monotonic decreasing and so the conventional tolerance limit is reached very quickly. This becomes clear when we consider the length of the straight ridge estimator $\hat{\beta}^* = (X'X + kI)^{-1}X'X\hat{\beta}$.

$$\phi(k) = \hat{\beta}^{*\prime}\hat{\beta}^* = \hat{\beta}'X'X(X'X+kI)^{-2}X'X\hat{\beta} \qquad (5.34)$$

$$= \sum_{i=1}^{n+1} \frac{\lambda_i^2\hat{\gamma}_i^2}{(\lambda_i+k)^2}$$

$$\frac{d\phi(k)}{dk} = -2\sum_{i=1}^{n+1} \frac{\lambda_i^2\hat{\gamma}_i^2}{(\lambda_i+k)^3} < 0 \; . \qquad (5.35)$$

where λ_i's are e-roots of $X'X$ and $\hat{\gamma} = G\hat{\beta}$.

Clearly, the length of $\hat{\beta}^*$ is decreasing monotonically as a function of k. The rate of decrease is faster at small values of k than at large values. The same applies to the length of $R\hat{\beta}^{(t)}$ as a function of k. It is therefore not surprising to find that the iterative ridge-type estimator converge quickly to the Almon Estimator under the usual tolerance limit for convergence. On the other hand, if iteration were applied to a generalised ridge estimator

$$\hat{\beta}^* = (X'X + GKG')^{-1}X'y \qquad (5.36)$$

where

$K = \text{diag}\{k_1 \ldots k_{n+1}\}$

and iteration begins with $k_i = \frac{\hat{\sigma}^2}{\hat{\gamma}_i^2}$

G = matrix of e-vectors of $X'X$

then it is possible that $\hat{\beta}^*$ will converge to a value different from zero. [See Hemmerle (1975)]. However, they are not very interesting estimators in this context because they will give irregular lag distributions. One way to avoid the convergence to the Almon estimator in the case of iterated Shiller estimator is to truncate the iterative process at a suitable point.

The Bayesian interpretation of the ridge type estimator has greatly enriched the class of alternative estimators suitable for the finite distributed lag model. We can now turn to some useful extensions of the technique in econometric analysis.

§5.3 Seasonal Variability

We consider the problem of estimating a finite distributed lag model with n lags corresponding to each season. Data are collected in s seasons and it is believed that the lag coefficients vary seasonally. The model can be written

$$y_t = \sum_{j=1}^{s} \sum_{i=1}^{n} \beta_{ij} D_{j,t-i} X_{t-i} + \varepsilon_t \qquad t = -n,\ldots,1,\ldots,T \tag{5.37}$$

where $D_1 \ldots D_s$ are seasonal dummy variables of the zero-one type. Under the restriction of a common distributed lag response for all s seasons the model reduces to

$$y_t = \sum_{i=1}^{n} \bar{\beta}_i X_{t-i} + \varepsilon_t \tag{5.38}$$

where

$$\beta_{i1} = \beta_{i2} = \ldots = \beta_{is} = \bar{\beta}_i .$$

Within each season, one can follow Shiller to impose a smoothness prior on the elements $\{\bar{\beta}_1 \ldots \bar{\beta}_n\}$. Gersovitz and MacKinnon (1978)

exploited the same idea by allowing for smoothness prior in the seasonal dimension.[1] In this section, the problem of estimating (5.37) is formulated in such a way that we can allow for 2-dimensional smoothness priors on the $\{\beta_{ij} | i = 1 \ldots n,\ j = 1 \ldots s\}$ and thus extend previously available results. We also consider estimation under the assumption of exchangeable priors on the $\{\beta_{ij}\}$ either in the seasonal dimension or in the lag dimension or simultaneously in both dimensions. Finally, the problem of estimation under a combination of exchangeable prior in the seasonal dimension and a smoothness prior in the lag dimension is considered. The total result is that we obtain a family of distributed lag estimators derived from a common Bayesian approach using the hierarchical prior model developed in Lindley and Smith (1972). The justification for considering "smooth seasonality" is particularly strong when monthly or even weekly data are being analysed.

Let us define the following matrices and vectors

$$\underset{(\tau\times s)}{Y} = [\underset{\sim}{y}_1 \ldots \underset{\sim}{y}_s] \qquad \underset{(\tau s\times 1)}{y} = \text{vec } Y$$

$$\underset{(\tau s\times ns)}{X} = \begin{bmatrix} X_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & X_s \end{bmatrix} \qquad \underset{(n\times s)}{B} = \begin{bmatrix} \beta_{11} & \cdots & \beta_{1s} \\ \vdots & & \\ \beta_{n1} & \cdots & \beta_{ns} \end{bmatrix}, \quad \underset{(ns=m)}{\beta} = \text{vec } B$$

$$\underset{(\tau\times s)}{E} = [\underset{\sim}{\varepsilon}_1 \ldots \underset{\sim}{\varepsilon}_s] \qquad \underset{(T\times 1)}{\varepsilon} = \text{vec } E\ .$$

The complete model with ns coefficients may be written in the stacked vector form

---

[1] As they did not explicitly concern themselves with a distributed lag model, these authors did not need to consider the possibility of simultaneous smoothing in the *seasonal* and the *lag* dimensions.

$$y = X\beta + \varepsilon \tag{5.39}$$

or

$$\underset{(\tau\times 1)}{y_{\underset{\sim}{g}}} = \underset{(\tau\times n)(n\times 1)}{X_g \beta_{\underset{\sim}{g}}} + \underset{(\tau\times 1)}{\varepsilon_{\underset{\sim}{g}}} \qquad (g = 1 \ldots s) . \tag{5.40}$$

We shall assume $\varepsilon \sim N(0,\sigma^2 I_T)$ $(T = \tau s)$. The following prior distributions are considered for the rows (within seasons) and columns (between seasons) of $\mathcal{B}$.

(i) Exchangeable priors between seasons (EBs), between elements within a season (EWs) and both (EBs×EWs).

(ii) Smoothness priors between seasons (SBs) (á lá Shiller), within season (SWs) (á lá Gersovitz and MacKinnon) and both (SBs×SWs).

(iii) Mixed priors of Exchangeability and Smoothness (EBs×SWs) or (SBs×EWs).

In dealing with the Exchangeability Between seasons (EBs), the set-up is as follows:-

$$\underset{(T\times 1)}{y} \sim N(X\beta,\sigma^2 I_T) \tag{5.41}$$

$$\underset{(m\times 1)}{\beta} \sim N(A\bar{\underset{\sim}{\beta}}, I_s \otimes \Sigma) \qquad A = \underset{\sim}{\ell}_s \otimes I_n \tag{5.42}$$

$\underset{\sim}{\ell}_s$ is an s-vector of units

Σ is $n\times n$ positive definite symmetric.

This is equivalent to saying that the seasonal variation of $\underset{\sim}{\beta}_g$ is random around the common location $(n\times 1)$ vector $\bar{\underset{\sim}{\beta}}$. If we have no information about $\bar{\underset{\sim}{\beta}}$ at all, we can assume a diffuse second stage prior to obtain the posterior means.

$$\underset{\sim}{\beta}^{*}_{g,EBs} = [\frac{1}{\sigma^2} X'_g X_g + \Sigma^{-1}]^{-1}[\frac{1}{\sigma^2} X'_g X_g \underset{\sim}{\hat{\beta}}_g + \Sigma^{-1}\underset{\sim}{\bar{\beta}}] \qquad g = 1 \ldots s \qquad (5.43)$$

where

$$\underset{\sim}{\bar{\beta}} = \sum_{g=1}^{s} W_g \underset{\sim}{\hat{\beta}}_g \quad \text{and} \quad \underset{\sim}{\hat{\beta}}_g = (X'_g X_g)^{-1} X'_g y_g \qquad (5.44)$$

$$W_g = \left[\sum_{j=1}^{s} \left(\frac{X'_j X_j}{\sigma^2} + \Sigma^{-1}\right)^{-1} \frac{X'_j X_j}{\sigma^2}\right]^{-1} \left(\frac{X'_g X_g}{\sigma^2} + \Sigma^{-1}\right)^{-1} \frac{X'_g X_g}{\sigma^2} . \qquad (5.45)$$

On the other hand, the notion of exchangeability within seasons (EWs) can be handled easily in the set-up

$$\underset{\sim}{y}_g \sim N(X_g \underset{\sim}{\beta}_g, \sigma^2 I_\tau) \qquad g = 1 \ldots s \qquad (5.46)$$

$$\underset{\sim}{\beta}_g \sim N(\underset{\sim}{\ell}_n \mu_g, \sigma_g^2 I_n) . \qquad (5.47)$$

With a diffuse prior on the scalar μ_g, we obtain the posterior mean in this case as

$$\underset{\sim}{\beta}^{*}_{g,EWs} = \{X'_g X_g + \frac{\sigma^2}{\sigma_g^2}(I_n - \frac{1}{n} J_n)\}^{-1} X'_g y_g \qquad (5.48)$$

where J_n is the $n \times n$ matrix whose every element is unity.

If the exchangeability assumption is applicable both between and within seasons (EBs×EWs), the additional restriction $\sigma_g^2 = \sigma_*^2$, all g, will be imposed and this leads to the posterior mean

$$\beta^{*}_{EBs\times EWs} = [X'X + \frac{\sigma^2}{\sigma_*^2}(I_m - \frac{1}{m} J_m)]^{-1} X'y \qquad m = ns . \qquad (5.49)$$

Note that (5.48) and (5.49) have the form of a ridge-type estimator. If we have a firm conviction that μ should be zero, then the diffuse prior at the second stage is unnecessary but (5.48) and (5.49) degenerate to the straight ridge estimator. Experience with using both the estimators

at (5.48) and the straight ridge estimators reveal that the shape of the estimated distribution of the lag weights is not compatible with our prior knowledge about the true distribution of lag weights [see Maddala (1977)]. Their usefulness in the estimation of distributed lag model seems to be limited.

On the other hand, applications of Shiller-type estimators for the simple one-dimensional distributed lag models show that the shape of the distribution of the estimated lag weights is more reasonable. This technique is now extended to the 2-dimensional distributed lag models to take account of seasonal variability of the lag weights.

Let us consider first the case of 2-dimensional smoothness between and within seasons [SBs×SWs] and specialise the results to the simple cases of 1-dimensional smoothness [SBs] and [SWs]. Suppose we wish to impose p-degree smoothness priors on the columns of $\mathcal{B}$ matrix and q-degree smoothness priors on the rows, $p < n$ and $q < s$. Let H_p and H_q denote, respectively, $(n\times p)$ and $(s\times q)$ Vandermonde matrices analogous to the one defined in (5.18). Let

$$\underset{(n\times s)}{\mathcal{B}} = \underset{(n\times p)}{H_p}\ \Gamma\ \underset{(q\times s)}{H_q'} + V \tag{5.50}$$

where Γ is $(p\times q)$ matrix of unknown constants and V is $(n\times s)$ matrix of random disturbances. Also

$$\beta = \text{vec}\,\mathcal{B} = [H_q \otimes H_p]\text{vec}\,\Gamma + \text{vec}\,V \tag{5.51}$$

i.e.

$$\underset{(m\times 1)}{\beta} = \underset{(m\times p)(q\times 1)}{H\gamma} + \underset{(m\ 1)}{v} \qquad \mathcal{E}(v) = 0 \qquad \mathcal{E}(vv') = \underset{(m\times m)}{\Omega} \qquad (m = ns)$$

The Bayes-Almon estimator, based on the assumption (5.51) with

$v \sim N_m(0,\Omega)$ and a second diffuse prior on γ, is

$$\beta^*_{BA} = \left\{ \frac{X'X}{\sigma^2} + \Omega^{-1} - \Omega^{-1}H(H'\Omega^{-1}H)^{-1}H'\Omega^{-1} \right\}^{-1} \frac{X'y}{\sigma^2} . \qquad (5.52)$$

Shiller's smoothness prior on β can be represented by

$$R\beta \sim N(0,\Sigma) \qquad (5.53)$$

where

$$\underset{(s-q)(n-p)\times m}{R} = [\ \underset{(s-q)\times s}{R_q} \otimes \underset{(n-p)\times n}{R_p}\ ]$$ a Kronecker product of

differencing matrices so that $R_qH_q = 0$ and $R_pH_p = 0$.

Combining (5.39) and (5.53), one obtains the Theil-Goldberger's mixed estimator

$$\beta^*_{SBs\times SWs} = [\frac{X'X}{\sigma^2} + R'\Sigma^{-1}R]^{-1} \frac{X'y}{\sigma^2} \qquad (5.54)$$

where $\Sigma = R\Omega R'$.

For the simple 1-dimensional case, one needs only either to replace H_p and R_p by I_n to obtain β^*_{SBs} (á lá Shiller) or to replace H_q and R_q by I_s to obtain β^*_{SWs} (á lá Gersovitz and MacKinnon).

Finally, the analysis involving mixed priors could be handled as follows. Suppose we have reason to believe that lag coefficients between seasons are 'similar' (EBs) and each seasonal set may be characterised by a smoothness assumption (SWs). Then, the specification is

$$\underset{\sim}{\beta}_g \sim N(\underset{\sim}{\bar{\beta}},\Omega) \qquad g = 1 \ldots s \qquad (5.55)$$

but

$$R_p\underset{\sim}{\bar{\beta}} \sim N(0,\Sigma) \qquad (5.56)$$

and this implies that one can apply the Shiller's estimator for each

season to get

$$\underset{\sim}{\beta}_g^* = (X_g'X_g/\sigma^2 + R_p\Sigma^{-1}R_p)^{-1}X_g'y_g/\sigma^2 \qquad g = 1 \ldots s. \tag{5.57}$$

However, this set-up has ignored the information that β_g $(g = 1 \ldots s)$ should be random around a common location. To incorporate this information formally, let us consider the hierarchical priors (we have assumed $\text{var}(y) = \sigma^2 I_T$ for simplicity but is otherwise not necessary)

$$\begin{aligned} y &\sim N(X\beta, \sigma^2 I_T) \\ \beta &\sim N(A_1\theta_1, C_1) \\ \theta_1 &\sim N(A_2\theta_2, C_2) \\ \theta_2 &\sim N(A_3\theta_3, C_3) \end{aligned} \tag{5.58}$$

which implies that the marginal prior distribution of β is

$$N(A_1A_2A_3\theta_3, C_1 + A_1C_2A_1' + A_1A_2C_3A_2'A_1') \ . \tag{5.59}$$

Let $B = C_1 + A_1C_2A_1'$, $E = A_1A_2$, then we can write the variance matrix of the marginal prior distribution of β as $B + EC_3E'$ whose inverse is

$$B^{-1} - B^{-1}E(E'B^{-1}E + C_3^{-1})^{-1}E'B^{-1} \tag{5.60}$$

and the posterior distribution of β is

$$N(Dd, D) \tag{5.61}$$

where

$$D^{-1} = \frac{1}{\sigma^2}X'X + (B + EC_3E')^{-1} \tag{5.62}$$

$$d = \frac{1}{\sigma^2}X'y + (B + EC_3E')^{-1}EA_3\theta_3 \ . \tag{5.63}$$

Suppose, now that the prior distribution of θ_2 is diffuse, i.e., $C_3^{-1} = 0$, then the posterior distribution of β can be shown to be

$$N(D_o d_o, D_o) \tag{5.64}$$

where

$$D_o^{-1} = \frac{1}{\sigma^2} X'X + B^{-1} - B^{-1}E(E'B^{-1}E)^{-1}E'B^{-1} \tag{5.65}$$

$$d_o = \frac{1}{\sigma^2} X'y = \frac{1}{\sigma^2} X'X\hat{\beta} \ . \tag{5.66}$$

For the mixed prior of EBs×SWs, we can specialise the above results by recognizing the appropriate matrices that define B and E in (5.65) and (5.66) as

$$A_1 = I_s \otimes I_n = I_m \qquad \theta_1 = \begin{pmatrix} \bar{\underset{\sim}{\beta}} \\ \vdots \\ \bar{\underset{\sim}{\beta}} \end{pmatrix} = \underset{\sim}{\ell}_s \otimes \bar{\underset{\sim}{\beta}} \qquad C_1 = I_s \otimes \Sigma$$

$$A_2 = \underset{\sim}{\ell}_s \otimes H_p \qquad C_2 = V$$

so that

$$B = C_1 + A_1 C_2 A_1' = I_s \otimes (\Sigma+V)$$

$$E = A_1 A_2 = (I_s \otimes I_n)(\underset{\sim}{\ell}_s \otimes H_p) = \underset{\sim}{\ell}_s \otimes H_p$$

$$(E'B^{-1}E)^{-1} = \frac{1}{s}\,(H_p'(\Sigma+V)^{-1}H_p)^{-1}$$

$$B^{-1}E(E'B^{-1}E)^{-1}E'B^{-1} = \frac{1}{s}\,(J_s \otimes K)$$

where

$$K = (\Sigma+V)^{-1}H_p[H_p'(\Sigma+V)^{-1}H_p]^{-1}H_p'(\Sigma+V)^{-1}$$

$$\frac{1}{\sigma^2} X'X + B^{-1} = \begin{pmatrix} \frac{X_1'X_1}{\sigma^2} + (\Sigma+V)^{-1} & & \bigcirc \\ & \ddots & \\ \bigcirc & & \frac{X_s'X_s}{\sigma^2} + (\Sigma+V)^{-1} \end{pmatrix}$$

$$= \begin{pmatrix} M_1 & & \bigcirc \\ & \ddots & \\ \bigcirc & & M_s \end{pmatrix}$$

where

$$M_g = \frac{X_g'X_g}{\sigma^2} + (\Sigma+V)^{-1} \qquad g = 1 \ldots s \ .$$

Then, the posterior mean in (5.64) of β satisfies the relation

$$D_o^{-1}\beta^* = d_o \tag{5.67}$$

i.e. from (5.65) and (5.66)

$$\begin{pmatrix} M_1\underset{\sim}{\beta}_1^* \\ \vdots \\ M_s\underset{\sim}{\beta}_s^* \end{pmatrix} - \frac{1}{s}\begin{pmatrix} K \\ \vdots \\ K \end{pmatrix} \sum_{g=1}^{s} \underset{\sim}{\beta}_g^* = \begin{pmatrix} \frac{1}{\sigma^2} X_1'y_1 \\ \vdots \\ \frac{1}{\sigma^2} X_s'y_s \end{pmatrix} .$$

Alternatively,

$$\begin{pmatrix} \beta_1^* \\ \vdots \\ \beta_s^* \end{pmatrix} - \frac{1}{s}\begin{pmatrix} M_1^{-1}K \\ \vdots \\ M_s^{-1}K \end{pmatrix} \sum_{g=1}^{s} \underset{\sim}{\beta}_g^* = \begin{pmatrix} \frac{1}{\sigma^2} M_1^{-1}X_1'y_1 \\ \vdots \\ \frac{1}{\sigma^2} M_s^{-1}X_s'y_s \end{pmatrix} . \tag{5.68}$$

Adding up these s equations to give

$$\sum_{g=1}^{s} \underset{\sim}{\beta}_g^* - \frac{1}{s}\left(\sum_{g=1}^{s} M_g^{-1}K \right)\left(\sum_{g=1}^{s} \underset{\sim}{\beta}_g^* \right) = \sum_{g=1}^{s} M_g^{-1} \frac{X_g'y_g}{\sigma^2}$$

i.e.

$$\left(I - \frac{1}{s} \sum_{i=1}^{s} M_i^{-1}K \right)\left(\frac{1}{s} \sum_{g=1}^{s} \underset{\sim}{\beta}_g^* \right) = \frac{1}{s} \sum_{g=1}^{s} M_g^{-1} \frac{X_g'X_g}{\sigma^2} \hat{\underset{\sim}{\beta}}_g \tag{5.69}$$

i.e.

$$\frac{1}{s} \sum_{g=1}^{s} \underset{\sim}{\beta}_g^* = \left(I - \frac{1}{s} \sum_{i=1}^{s} M_i^{-1}K \right)^{-1} \frac{1}{s} \sum_{g=1}^{s} M_g^{-1} \frac{X_g'X_g}{\sigma^2} \hat{\underset{\sim}{\beta}}_g \ . \tag{5.70}$$

Hence, plugging (5.70) into (5.68), the appropriate Lindley-Smith estimator is

$$\underset{\sim}{\beta}_j^* = \frac{1}{\sigma^2} M_j^{-1} X_j' y_j + M_j^{-1} K \left(I - \frac{1}{s} \sum_{i=1}^{s} M_i^{-1} K \right)^{-1} \frac{1}{s} \sum_{g=1}^{s} M_g^{-1} \frac{X_g' X_g}{\sigma^2} \underset{\sim}{\hat{\beta}}_g$$

$$= M_j^{-1} \left\{ \frac{1}{\sigma^2} X_j' y_j + K \underset{\sim}{\bar{\beta}} \right\} \qquad j = 1 \ldots s \tag{5.71}$$

where

$$\underset{\sim}{\bar{\beta}} = \frac{1}{s} \sum_{g=1}^{s} \underset{\sim}{\beta}_g^* = \left[\sum_{i=1}^{s} (I - M_i^{-1} K) \right]^{-1} \sum_{g=1}^{s} M_g^{-1} \frac{X_g' X_g}{\sigma^2} \underset{\sim}{\hat{\beta}}_g \tag{5.72}$$

$$= \sum_{g=1}^{s} W_g \underset{\sim}{\hat{\beta}}_g$$

$$W_g = \left[\sum_{i=1}^{s} M_i^{-1} (M_i - K) \right]^{-1} M_g^{-1} \frac{X_g' X_g}{\sigma^2} \qquad g = 1 \ldots s \tag{5.73}$$

$$\hat{\beta}_g = (X_g' X_g)^{-1} X_g' y_g \, .$$

Note. If $H_p \equiv I_n$, then $K \equiv (\Sigma + V)^{-1}$. It is easy to see that

$$M_j - K = \frac{X_j' X_j}{\sigma^2}$$

so that (5.73) becomes

$$W_g = \left[\sum_{i=1}^{s} M_j^{-1} \frac{X_j' X_j}{\sigma^2} \right] M_g^{-1} \frac{X_g' X_g}{\sigma^2} \tag{5.74}$$

which agrees with the result of Lindley-Smith (1972) for exchangeability between seasons alone without the complicity of smoothness within season.

The reverse case of smoothness across seasons but exchangeability within season can also be handled in a similar fashion.

In summary, the hierarchical prior model of Lindley and Smith embraces a number of cases considered in the literature. The main novel

feature of this section is to establish the connections between various estimators by combining the assumptions of smoothness and exchangeability priors. Our discussion applies to the case of known variances only. To extend to the case of unknown variances we may follow Lindley and Smith and specify additional prior distributions on all variance parameters. Needless to say the computational problems involved are not trivial.

§5.4 Estimation in Frequency Domain

A number of "good" estimators in the time domain have been suggested and extended for econometric analysis. Unfortunately, the above analysis has been based on the simplifying assumption that the disturbance ε_t is Gaussian noise. In theory, the finite lag model (5.15) is only an approximation to the infinite lag model (5.1). If the tail of the lag distribution is not negligibly small, then the catch-all disturbance term will be far from being a Gaussian noise. We are bound to confront a serial correlation problem and it is well known that OLS is inefficient in this situation. Specifically, if the covariance matrix of the disturbance vector ε is Ω, then the efficiency of OLS is bounded below by

$$\prod_{i=1}^{n^*} \left\{ \frac{4d_i d_{T+1-i}}{(d_i + d_{T+1-i})^2} \right\} \tag{5.75}$$

where

$$n^* = \min\{n, T-n\}$$

$$d_1 < d_2 < \ldots < d_T \text{ are e-roots of } \Omega.$$

[See Knott (1975)].

If ε_t is known to follow some finite parametric processes, which can be filtered to become white noise, then OLS will remain efficient

when applied to the pre-filtered data. In practice, the process of ε_t is unknown and the usual approach is to approximate it by a low order autoregressive-moving average (ARMA) process [see Nicholls et al (1975)]. Engle (1974) called the generalised least square procedure taking account of this approximated error process the truncated Aitken estimator since the true error process might be of infinite order. Of course, if the truncation happens to be true, then the truncated Aitken estimator is asymptotically efficient. Ironically, although OLS is a special form of truncated Aitken estimator, namely, when all the estimated serial correlation are zero, it is sometimes found that OLS performs better than some ill-determined truncated Aitken estimators. It was established by Engle (1974) that the condition for the truncated Aitken estimator to be more efficient than OLS is that the true spectrum of the disturbances and the estimated one must *never* have slopes of opposite sign. Since the true disturbance process is likely to have many changes in slope, only a high order process is possible to mirror the changes in slopes of the true spectrum. The results of Engle suggest that the usual first order Markov approximation to a generally high order process is not appropriate. This agrees with Malinvaud (1970)'s rule of thumb that only if the first order serial correlation "considerably exceeds" 0.5 should one look for an estimator other than OLS. Of course, if one is interested in drawing inference from the estimates, then the t-ratios or F-statistics based on OLS is misleading since one typically would not be able to use the true variance matrix of OLS. Nicholls and Pagan (1977) showed that the discrepancy between OLS and the best linear unbiased estimator depends on the type of the error process. They also advocated the iterative Aitken procedure as suggested in Wallis (1972).

The difficulties of specifying a valid parametric ARMA process for the disturbance term could be avoided if we estimate the distributed lag

model in the frequency domain. The reason is that it allows non-parametric identification of the power spectrum of the disturbance and other variables in the model. More importantly, estimation of the distributed lag model is particularly simple in the frequency domain, which will become clear as we proceed.

Consider again the finite distributed lag model in matrix notation as

$$\underset{(T\times 1)}{y} = \underset{(T\times n)(n\times 1)}{X\beta} + \underset{(T\times 1)}{\varepsilon} \qquad \begin{matrix} \mathcal{E}(\varepsilon) = 0 \\ \mathcal{E}(\varepsilon\varepsilon') = \Omega \end{matrix} \tag{5.76}$$

The observations are *standardised* (dividing the deviates from the sample means by its root sum of squares) so that $X'X$ is the matrix of sample autocorrelation functions

$$X'X = \begin{bmatrix} r(0) & \cdots & r(n-1) \\ \vdots & & \\ r(n-1) & & r(0) \end{bmatrix}, \quad r(0) \equiv 1 \, . \tag{5.77}$$

If x_t is a weakly stationary time series, then the Toeplitz matrix (5.77) has e-roots approximately equal to [see Amemiya and Fuller (1967)]

$$\lambda_j = 2\pi f_x(w_j) \qquad j = 1 \ldots n \tag{5.78}$$

where $f_x(w_j)$ is the spectral density function of x-process at frequency $w_j = \frac{2\pi j}{n}$.

i.e. $$f_x(w_j) \overset{\text{def}}{=\!=} \frac{1}{2\pi} \sum_{\ell=o}^{n-1} r(k) e^{ikw_j}$$

and has the corresponding e-vectors

$$\underset{\sim}{h}_j = \frac{1}{\sqrt{n}} \begin{pmatrix} 1 \\ e^{iw_j} \\ \vdots \\ e^{i(n-1)w_j} \end{pmatrix} \qquad j = 1 \ldots n \tag{5.79}$$

where $i = \sqrt{-1}$.

It is well known that the e-roots of a symmetric matrix whose elements are real numbers are themselves real numbers. For a stationary circular process, $r(\ell) = r(n-\ell)$ and $e^{i2\pi\ell} = \cos(2\pi\ell) \equiv 1$, we can write, assuming n is an odd integer,

$$\lambda_j = r(0) + \sum_{\ell=1}^{[\frac{n}{2}]} r(\ell) \left[e^{\frac{2\pi i \ell j}{n}} + e^{\frac{2\pi i (n-\ell) j}{n}} \right] \tag{5.80}$$

where $[x]$ = the largest integer below x

$$= r(0) + 2 \sum_{\ell=1}^{[\frac{n}{2}]} r(\ell) \cos \ell w_j \qquad j = 1 \ldots n \, .$$

Similarly,

$$\lambda_{n-j} = r(0) + 2 \sum_{\ell=1}^{[\frac{n}{2}]} r(\ell) \cos \frac{2\pi(n-\ell)j}{n} = \lambda_j \, . \tag{5.81}$$

This implies that the real e-roots occur in $\frac{n}{2}$ equal pairs. However, the e-vectors $\underset{\sim}{h}_j$ are complex (in pairs of conjugates corresponding to the double e-roots). One can construct real vectors for the odd and even numbered columns respectively as,

$$\underset{\sim}{F}_{2j-1} = \frac{1}{\sqrt{2}} (\underset{\sim}{h}_j + \underset{\sim}{h}_{n-j}) = \left[\sqrt{\frac{2}{n}} \cos \frac{2\pi j \ell}{n} \, , \quad \ell = 0, \ldots, n-1 \right]$$

$$\underset{\sim}{F}_{2j} = \frac{1}{i\sqrt{2}} (\underset{\sim}{h}_j - \underset{\sim}{h}_{n-j}) = \left[\sqrt{\frac{2}{n}} \sin \frac{2\pi j \ell}{n} \, , \quad \ell = 0, \ldots, n-1 \right]$$

$$j = 1, 2, \ldots, \frac{n}{2} \, .$$

The $(n \times n)$ matrix F whose columns consist of these vectors is orthogonal and is called the real Fourier Matrix, which is an approximation to the unitary matrix whose columns are the complex e-vectors $\underset{\sim}{h}_j$.

Exploiting the special characteristic of the matrix of $X'X$ in a distributed lag model, Trivedi (1978) exhibited Bock (1975)'s conditions for the James-Stein estimator to MSE-dominate OLS in terms of theoretical spectral density of the explanatory variable. He also established the various regions of superiority among the OLS, Stein-estimator, Preliminary Test Estimator and Restricted Least Square estimator for a number of stationary processes. For instance, when the x_t process is AR1, namely,

$$x_t = \rho_1 x_{t-1} + v_t \qquad \mathcal{E}(v_t^2) = \sigma_v^2$$

the James-Stein estimator will dominate OLS when

$$\frac{m(1+\rho_1^2)}{(1+\rho_1)^2} > 2 \tag{5.82}$$

where m = the number of restrictions which is assumed to exceed 2 while the Restricted Least Square estimator will dominate OLS if

$$\frac{\sigma_v^2}{2\sigma^2 T} \sum_{i=1}^{m} \beta_i^2 (1+\rho_1^2) \leq \frac{m}{2} \frac{(1+\rho_1^2)}{(1-\rho_1)^2} \tag{5.83}$$

and finally, the Preliminary Test Estimator will dominate the Restricted Least Square estimator if

$$\frac{\sigma_v^2}{2\sigma^2 T} \sum_{i=1}^{m} \beta_i^2 \geq \frac{m}{2(1+\rho_1)^2} \,. \tag{5.84}$$

The size of ρ_1 obviously plays a crucial role to determine the regions of superiority.

In this section, we focus our attention on the development of the Spectral Ridge-type Estimator (SRE) and will take special note of the serially correlation problem of the disturbance term which has been ignored in previous analyses. Consider the Aitken Estimator (Generalised Least Squares)

$$\tilde{\beta}_{GLS} = \left(\frac{X'\Omega^{-1}X}{T} \right)^{-1} \left(\frac{X'\Omega^{-1}y}{T} \right) = Q^{-1}\underset{\sim}{g} \; . \tag{5.85}$$

For T large and ε_t being a stationary process, Ω can be approximately diagonalised by the $(T\times T)$ unitary matrix H such that

$$H\Omega H^* \simeq D \simeq \text{diag}\{2\pi f_\varepsilon(w_j)\} \qquad j = 1 \ldots T \tag{5.86}$$

(for the sake of exposition, we have used the complex matrix).

Since $H^{-1} = H^*$, we have $(H\Omega H^*)^{-1} = H^{*-1}\Omega^{-1}H^{-1} = H\Omega^{-1}H^*$. Hence, Q can be written as

$$Q = \frac{X'\Omega^{-1}X}{T} = \frac{1}{T} X'H^*D^{-1}HX \tag{5.87}$$

$$Q_{\ell j} = \frac{1}{T} \sum_{g_1=1}^{T} \sum_{g_2=1}^{T} \sum_{g_3=1}^{T} \sum_{g_4=1}^{T} \frac{x_{g_1-\ell+1} e^{-i\lambda_{g_2} g_1} \delta_{g_1 g_3} e^{i\lambda_{g_3} g_4} x_{g_4-j+1}}{d_{g_3}} \tag{5.88}$$

where

$$\delta_{g_1 g_3} = 1 \quad \text{if} \quad g_1 = g_3$$

$$= 0 \quad \text{otherwise}$$

$$= \frac{1}{T} \sum_{g=1}^{T} \frac{1}{f_\varepsilon(w_g)} \left\{ \frac{1}{2\pi} \sum_{\tau=-T+1}^{T-1} \left(1 - \frac{|\tau|}{T}\right) r(\tau+\ell-j) e^{i\tau w_g} \right\} . \tag{5.89}$$

If we partition all the T frequency points into N frequency bands in such a way that $N \to \infty$ as $T \to \infty$ but $\frac{N}{T} \to 0$, then the

asymptotic value of $Q_{\ell j}$ remains unchanged (see Fishman (1969)) and the term inside curly bracket with T replaced by N becomes a consistent estimator of $f_x(w_g)e^{-i(\ell-j)w_g}$.

Thus for large T, we have approximately

$$Q_{\ell j} = \frac{1}{N}\sum_{g=1}^{N} f_{\varepsilon}^{-1}(w_g)\hat{f}_x(w_g)e^{-i(\ell-j)w_g} \qquad \ell = 1 \ldots n \qquad j = 1 \ldots n \,. \tag{5.90}$$

Similarly, the j'th element of q is

$$q_j = \frac{1}{N}\sum_{g=1}^{N} f_{\varepsilon}^{-1}(w_g)\hat{f}_{yx}(w_g)e^{-ijw_g} \qquad j = 1 \ldots n \,. \tag{5.91}$$

where $f_{yx}(\cdot)$ is the cross spectrum of x, y.

If N is even, and define $n = \frac{N}{2}$, then the special distributed lag model

$$y_t = \sum_{j=-n+1}^{n} \beta_j X_{t-j} + \varepsilon_t \tag{5.92}$$

has an (N×N) matrix X'X. The matrix Q defined in (5.90) may be written in the approximate form

$$Q = H^* \operatorname{diag}\left\{\frac{\hat{f}_x(w_g)}{f_{\varepsilon}(w_g)}\right\} H \tag{5.93}$$

and, the vector q defined in (5.91) may be written as

$$\underset{\sim}{q} = H^* \begin{pmatrix} \dfrac{\hat{f}_{yx}(w_1)}{f_{\varepsilon}(w_1)} \\ \vdots \\ \dfrac{\hat{f}_{yx}(w_N)}{f_{\varepsilon}(w_N)} \end{pmatrix} . \tag{5.94}$$

Thus, the Aitken Estimator for the special model (5.92) is,

$$\tilde{\beta}_{GLS} = Q^{-1}q = H^* \text{ diag } \left[\frac{\hat{f}_\varepsilon(w_g)}{\hat{f}_x(w_g)} \right] HH^* \begin{pmatrix} \frac{\hat{f}_{yx}(w_1)}{\hat{f}_\varepsilon(w_1)} \\ \vdots \\ \frac{\hat{f}_{yx}(w_N)}{\hat{f}_\varepsilon(w_N)} \end{pmatrix}$$

$$= H^* \begin{pmatrix} \frac{\hat{f}_{yx}(w_1)}{\hat{f}_x(w_1)} \\ \vdots \\ \frac{\hat{f}_{yx}(w_N)}{\hat{f}_x(w_N)} \end{pmatrix} \tag{5.95}$$

i.e.

$$\tilde{\beta}_j = \frac{1}{N} \sum_{\ell=-n+1}^{n} \frac{\hat{f}_{yx}(\phi_\ell)}{\hat{f}_x(\phi_\ell)} e^{-ij\phi_\ell} \qquad j = -n+1,\ldots,n \tag{5.96}$$

where

$$\phi_\ell = w_{\ell+n} = \frac{\pi j}{n}$$

which is simply the inverse-Fourier Transform of the sequence

$$\left\{ \frac{\hat{f}_{yx}(\phi_\ell)}{\hat{f}_x(\phi_\ell)} , \ \ell = -n+1,\ldots,n \right\}.$$

The estimator (5.96) has the same form as the Hannan's Simpler Estimator (HSE) which is consistent and efficient if (5.92) is the correct model. However, in general (5.92) is not the true model, Hannan (1965) justified the estimator as being "**nearly**" efficient if the ratio $\frac{\hat{f}_x(w)}{f_\varepsilon(w)}$ is nearly constant for all frequencies w.

Since the covariance matrix is diagonal in this special case, Hannan's simpler estimator is particularly useful as an exploratory device to identify the lag-length with the minimum amount of computation.

Cargill and Meyer (1974)'s simulation compared OLS, HSE and the Almon Estimator and they concluded that the HSE is generally the second best choice in terms of small bias and robustness with respect to misspecification. Further evidences of good performance were provided by Doran (1974) and Engle and Gardner (1976) who suggested that it pays to switch to the Hannan's estimator at sample size 100.

If the model (5.92) is assumed to be approximately correct, Doran (1976) extended the idea of the classical principal component analysis in the frequency domain. Analogous to the truncation of small e-roots in the principal component analysis, Doran eliminates frequency bands corresponding to small signal-noise ratios to give the Spectral Principal Component Estimator (SPCE). It is well known that, in time domain, the truncated Principal Component estimator does reduce the asymptotic variance of the Aitken estimator but is asymptotically biased and therefore inconsistent. The same applies to Doran's SPCE in the frequency domain. Doran performed some simulation experiments and found that in cases where the signal-noise ratio is far from being constant the SPCE could have substantial gain in precision over the HSE.

We have already studied the properties of the ridge-type estimators in the time domain, it is tempting to obtain equivalent estimators in the frequency domain. Some effort in this direction was made by Phillips (1962) and Hunt (1970) in obtaining numerical solution of some convolution-type integral equations. Unfortunately, by assuming a scalar covariance matrix, they failed to exploit the ability of the spectral analysis in obtaining consistent estimates of the error spectrum, which was our primary motivation to leave the time domain. We remedy this defect below and specifically relate the Spectral Ridge Estimator (SRE) to the other estimators for estimating the finite distributed lag model.

The ridge-type estimation under smoothness prior á lá Shiller, taking account of serially correlated disturbances in the time domain is

$$\tilde{\beta}_S = (X'\Omega^{-1}X + kR'R)^{-1}X'\Omega^{-1}y \ . \tag{5.97}$$

Note that $X'\Omega^{-1}X$ is a Toeplitz matrix which can be diagonalised by the unitary matrix H. On the other hand, the matrix R in the Shiller estimator is a matrix whose elements are coefficients of products of p forward differencing operators $\Delta = (1-L)$. For instance $p = 2$, gives

$$R = \begin{bmatrix} 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \end{bmatrix} \tag{5.98}$$

which is also a Toeplitz matrix and hence can be diagonalised by the same unitary matrix.

Writing the matrices in the frequency domain for the special model (5.92) as

$$X'\Omega^{-1}X = H^{*}\text{diag}\left\{\frac{\hat{f}_x(w_g)}{f_\varepsilon(w_g)}\right\}H \tag{5.99}$$

and

$$R'R = H^{*}\text{diag}\{\hat{f}_R(w_g)\}H \tag{5.100}$$

so that we have

$$(X'\Omega^{-1}X + kR'R) = H^{*}\left[\text{diag}\left\{\frac{\hat{f}_x(w_g)}{f_\varepsilon(w_g)}\right\} + k\,\text{diag}\{\hat{f}_R(w_g)\}\right]H \ . \tag{5.101}$$

This combines with the vector q defined in (5.94) to give the Spectral Ridge-type Estimator (SRE),

$$\tilde{\beta}^*_{SRE} = H^* \begin{bmatrix} \left[\frac{f_x(w_1)}{f_\varepsilon(w_1)} + kf_R(w_1) \right]^{-1} \frac{\hat{f}_{yx}(w_1)}{f_\varepsilon(w_1)} \\ \vdots \\ \left[\frac{\hat{f}_x(w_N)}{f_\varepsilon(w_N)} + kf_R(w_N) \right]^{-1} \frac{f_{yx}(w_N)}{f_\varepsilon(w_N)} \end{bmatrix} . \qquad (5.102)$$

The g'th element of $H\tilde{\beta}^*_{SRE}$ is given by

$$b(w_g) = \frac{\hat{f}_{yx}(w_g)}{\hat{f}_x(w_g) + kf_\varepsilon(w_g)\hat{f}_R(w_g)} \qquad g = 1 \ldots N \qquad (5.103)$$

and $\tilde{\beta}^*_{SRE}$ is seen to be the inverse Fourier transform of the sequence $\{b(w_g): g = 1 \ldots N\}$ in (5.103). $\hat{f}_R(w_g)$ and $\hat{f}_x(w_g)$ can be obtained by using the Fourier Transforms of the columns of R and columns of X respectively, whereas $f_\varepsilon(w_g)$ can be estimated from the transforms of OLS residuals. The ridge parameter k can be obtained in ways analogous to their counterparts in the time domain. Note that when $k = 0$, $\tilde{\beta}^*_{SRE}$ in (5.102) becomes the Hannan's Simpler Estimator $\tilde{\beta}_{GLS}$ in (5.95). On the other hand, if some k equal 0 and some are set at ∞, then $\tilde{\beta}^*_{SRE}$ becomes Doran's Spectral Principal Component Estimator. The advantage of using the estimator $\tilde{\beta}^*_{SRE}$ is that it allows for some compromise between the two extremes of Hannan's and Doran's procedure.

However, these three frequency domain methods are simply quick and dirty methods for obtaining a rough estimates of the lag weights. Modern computer's capacity proves that these techniques are not necessary as more accurate estimates could be obtained at only slightly higher cost. A direct calculation of the Hannan's efficient estimator (5.84) can be based on Q and q whose elements are defined in (5.90) and (5.91) respectively. This method is approximately equivalent to first applying a finite Fourier transform of the X matrix and the y vector,

(the covariance matrix of the transformed disturbance term is given approximately by the diagonal matrix whose T elements are spectral densities of the disturbance process at T different frequency points) and then applying Aitken's generalised least squares. Thus, it is now a simple matter to extend the idea of Shiller's smoothness prior in the frequency domain, which can be handled as a Theil-Goldberger mixed estimator for the model

$$\begin{aligned} y^* &= X^*\beta + \varepsilon^* \qquad \varepsilon^* \sim N(0,I) \\ 0 &= R\beta + v \qquad\quad v \sim N(0,\tfrac{1}{k} I) \end{aligned} \tag{5.104}$$

where $y^* = \Lambda^{-\frac{1}{2}}F'y \qquad X^* = \Lambda^{-\frac{1}{2}}F'X$

$\Lambda = \text{diag}\{f_\varepsilon(w_1) \ \ldots \ f_\varepsilon(w_T)\} = F'\Omega F$

Ω = the covariance matrix of the original disturbance

R = the appropriate differencing matrix.

The F matrix, which is orthogonal, is given as

$$F = \sqrt{\frac{2}{T}} \begin{bmatrix} \frac{1}{\sqrt{2}} & \cos\frac{2\pi}{T} & \sin\frac{2\pi}{T} & \cos 2(\frac{2\pi}{T}) & \ldots\ldots & \sin(\frac{T}{2}-1)(\frac{2\pi}{T}) & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \cos 2(\frac{2\pi}{T}) & \sin 2(\frac{2\pi}{T}) & \cos 4(\frac{2\pi}{T}) & & \sin(\frac{T}{2}-1)2\frac{2\pi}{T} & \frac{1}{\sqrt{2}} \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ \frac{1}{\sqrt{2}} & 1 & 0 & 1 & & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}$$

if T is even

and

$$F = \sqrt{\frac{2}{T}} \begin{bmatrix} \frac{1}{\sqrt{2}} & \cos\frac{2\pi}{T} & \sin\frac{2\pi}{T} & \cos 2\left(\frac{2\pi}{T}\right) & \cdots\cdots & \sin\left(\frac{T}{2}-1\right)\frac{2\pi}{T} \\ \frac{1}{\sqrt{2}} & \cos 2\left(\frac{2\pi}{T}\right) & \sin 2\left(\frac{2\pi}{T}\right) & \cos 4\left(\frac{2\pi}{T}\right) & & \sin 2\left(\frac{T}{2}-1\right)\frac{2\pi}{T} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ \frac{1}{\sqrt{2}} & 1 & 0 & 1 & & 0 \end{bmatrix}$$

if T is odd.

The 'efficient' spectral Ridge-type estimator is given by

$$\tilde{\beta}^*_{SRE} = (X^{*\prime}X^* + kR'R)^{-1}X^{*\prime}y^* \tag{5.105}$$

as opposed to the 'efficient' Hannan's estimator

$$\tilde{\beta}^*_{HE} = (X^{*\prime}X^*)^{-1}X^{*\prime}y^* \tag{5.106}$$

These estimators are valid for all general linear model including (5.92) as a special case.

§5.5 Performance of Spectral Ridge-type Estimator

There is a general reluctance among economists to consider spectral techniques. The reasons are notably:

(i) The complication of the spectral theory which involves an understanding of complex number.

(ii) The need of large amount of data because the statistical results in the frequency domain are asymptotic.

(iii) For better estimates of spectral densities, judgemental parameters such as spectral windows and truncation points are required.

(iv) The gain in efficiency over comparable time domain techniques are rather trivial.

(v) Consume enormous computer time.

Some of these criticisms are still relevant to the spectral ridge-type estimator but some are of only supervisial validity. Since spectral regression estimates are based on linear combinations of estimates of spectral densities, the inconsistency involved for estimating individual ordinates of the spectrum at particular frequency points does not affect the consistency of the regression estimates. The consumption of computer time for the various types of spectral estimates considered in the section is within tolerable limits and sometimes it may even be less than the time domain approach since the spectral densities are usually by-products of a preliminary analysis of economic time series. With the development of the fast Fourier Transform algorithm, time needed to obtain spectral estimates are actually substantially less than the corresponding time domain procedures. To illustrate the performance of the Spectral ridge-type estimator, we have used Almon's quarterly data on capital expenditure (y) and appropriations (x). These data are reproduced in Maddala (1977, p.370) and consist of 60 observations. To facilitate comparison with existing results, we have considered the lag distribution up to x_{t-8} only, namely,

$$y_t = \alpha + \beta_o x_t + \beta_1 x_{t-1} + \ldots + \beta_8 x_{t-8} + \varepsilon_t \qquad (5.107)$$

The model is re-estimated by six different procedures based on an effective sample size of 52. Among the six, there are four time domain procedures viz., OLS($\hat{\beta}$), truncated principal components ($\tilde{\beta}_{PC}$), Almon's restricted least squares ($\tilde{\beta}_A$) and Shiller's estimator ($\tilde{\beta}_S$) and two frequency domain procedures viz. Hannan's efficient estimator ($\tilde{\beta}_{HE}$)

and the efficient Spectral Ridge Estimator ($\tilde{\beta}_{SRE}$). The estimated lag weights are given as follows:-

$$\hat{\beta}'_{OLS} = \{.063, .087, .227, .186, .132, .018, .138, .066, .052\}$$

$R^2 = .99$ DW = .35 sample size = 52

$$\tilde{\beta}'_{PC} = \{.100, .119, .139, .148, .144, .126, .096, .063, .030\}$$

assigned rank = 3 (95% of total variation)

$$\tilde{\beta}'_{A} = \{.091, .121, .140, .148, .144, .130, .103, .066, .018\}$$

order of polynomial = 2

$$\tilde{\beta}'_{S} = \{.080, .122, .156, .165, .145, .109, .073, .055, .065\}$$

k = .240 order of differencing = 3

$$\tilde{\beta}'_{HE} = \{.158, .092, .260, .141, .130, -.069, .115, -.003, .142\}$$

$R^2 = .999$ DW = 2.08

$$\tilde{\beta}'_{SRE} = \{.167, .124, .138, .131, .117, .099, .077, .077, .092\}$$

k = .27 order of differencing = 3

From the OLS result, a saw-tooth lag distribution is evident and this is possibly due to strong collinearity among lagged x-variables. The Durbin-Watson statistic (DW = .35) indicates that a positively autocorrelated error is impairing the efficiency of OLS. The flat pattern of the truncated Principal Component estimates is no surprise as it can be shown that, for positively correlated x-variables, dropping of e-vectors corresponding to small e-roots is equivalent to imposing equality constraints among the lag weights.[2] The estimates provided by the Almon's and Shiller's procedures need no further explanation. Although the Hannan's efficient estimator has successfully removed the

---

[2] I owe this point to Dr S. Yeo of London School of Economics through private communication.

problem of serial correlation (Durbin-Watson statistic equals 2.08), it is unsatisfactory because of the erratic lag weights resulting from worsened multicollinearity among the transformed variables. The efficient Spectral Ridge estimator, on the other hand, provides sensible results.

Since the true lag weights are unknown, there is no way of giving an objective assessment of the performance of the six estimators. One thing that is more certain is that estimates based on a large sample of observation is generally more accurate than those based on a smaller sample. In order to compare the various estimators, we have re-estimated the same model using a reduced sample (size = 32). The total mean square error of each estimator are then calculated treating each of the above estimates in turn as if it were the true weights. Each estimator is expected to outperform the others when its estimates in the larger sample were used as the bench-mark. The estimated lag weights based on the smaller sample are as follows:-

$$\hat{\beta}'_{OLS} = \{-.004, .128, .169, .268, .012, .120, .043, .157, -.028\}$$

$R^2 = .92$ DW = .56 sample size = 32

$$\tilde{\beta}'_{PC} = \{-.014, .095, .175, .182, .145, .093, .055, .053, .058\}$$

assigned rank = 4 (95% of total variation)

$$\tilde{\beta}'_{A} = \{.078, .111, .132, .143, .142, .131, .108, .075, .030\}$$

order of polynomial = 2

$$\tilde{\beta}'_{S} = \{.028, .084, .150, .170, .143, .107, .083, .057, .028\}$$

k = .12 order of differencing = 3

$$\tilde{\beta}'_{HE} = \{.088, .049, .196, .222, .134, .030, .101, .072, .122\}$$

$R^2 = .999$ DW = 2.11

$$\tilde{\beta}'_{SRE} = \{.087, .117, .148, .132, .134, .108, .095, .093, .096\}$$

k = .25 order of differencing = 3.

The total mean square error of each estimator are listed in Table 5.5.1.

Table 5.5.1

Comparison of Estimators for a Distributed Lag Model

Bench-mark / Competitors	OLS	PC	Almon	Shiller	Hannan	Spectral Ridge	Total
$\hat{\beta}_{OLS}$	.076	.085	.071	.069	.184	.082	.557
$\tilde{\beta}_{PC}$	.022	.020	.020	.011	.079	.022	.174
$\tilde{\beta}_{A}$	.026	.001	.001	.005	.082	.008	.123
$\tilde{\beta}_{S}$	.019	.008	.007	.006	.079	.015	.134
$\tilde{\beta}_{HE}$	.011	.032	.035	.021	.033	.024	.156
$\tilde{\beta}_{SRE}$	.023	.006	.008	.004	.061	.002	.104

The minimum within each column is *underlined*.

From Table 5.5.1, it is clear that the Spectral Ridge being the most stable, is the winner overall. Although the Hannan's Efficient estimator, $\tilde{\beta}_{HE}$ is superior to all others in its own regime, yet it is badly beaten in some alien regimes. Note that the stability of ridge-type technique in the frequency domain has resulted in a much smaller total TMSE of $\tilde{\beta}_{SRE}$ than that of $\tilde{\beta}_{HE}$. It is clear that it pays to correct for serial correlation since $\tilde{\beta}_{HE}$ and $\tilde{\beta}_{SRE}$ outperform $\hat{\beta}_{OLS}$ and $\tilde{\beta}_{S}$ under the regimes of OLS and Shiller respectively. The fact that $\tilde{\beta}_{S}$ manages to have the smaller total is due to the robustness of the estimator. It will be shown in the next chapter that misspecification on the error structure will not seriously affect the performance of a ridge-type estimator.

In view of the relative accuracy and stability, a ridge-type procedure either in the time domain or in the frequency domain is

recommended. Of course, our result provides only one single illustration of the techniques. More extensive studies and applications are needed to gain confidence. We have not attempted a Monte Carlo study here because we believe that it would be too specific to a particular hypothetical error structure and a set of parameter values unless a really large scale and expensive experiment is undertaken.

CHAPTER 6

"IMPROVED" ESTIMATORS IN ECONOMETRICS II: SIMULTANEOUS EQUATION MODELS

§6.1 The Undersized Sample Problem

Consider the problem of estimation in a linear system of G simultaneous structural equations whose data matrix of T observations on each variable can be written as

$$YB + X\Gamma = U \tag{6.1}$$

where Y is $(T\times G)$ matrix of endogenous variables.

X is $(T\times K)$ matrix of predetermined variables.

U is $(T\times G)$ matrix of structural disturbances each row of which is G-dimensional multivariate normal with mean zero and covariance matrix Σ.

The typical structural equation consists of $G_1 + K_1 = n$ explanatory variables $Z_1 = (Y_1, X_1)$ where Y_1 is $(T\times G_1)$ submatrix of Y and X_1 is $(T\times K_1)$ submatrix of X, [We assume $G = G_1 + G_2 + 1$, $K = K_1 + K_2$] namely,

$$\underset{(T\times 1)}{\underset{\sim}{y}} = Y_1\underset{\sim}{\beta} + X_1\underset{\sim}{\gamma} + \underset{\sim}{u}\sigma = Z_1\underset{\sim}{\delta} + \underset{\sim}{u}\sigma \tag{6.2}$$

where $\underset{\sim}{\delta}' = (\underset{\sim}{\beta}' \quad \underset{\sim}{\gamma}')$ $\underset{\sim}{u}$ has mean zero and covariance I_n

σ is a small positive scalar. The smaller the σ, the better is the fit of the model.

Assuming that the system is complete and B is $(G\times G)$ nonsingular, each column of Y can then be "explained" by a linear combination of the

columns of X. Thus, we have the reduced form of the model

$$\underset{(T\times G)}{Y} = \underset{(T\times K)(K\times G)}{X\Pi} + V\sigma \tag{6.3}$$

where Π is the $(K\times G)$ matrix of reduced form coefficients. B and Γ satisfy the set of linear restrictions

$$B\Pi + \Gamma \equiv 0 \ . \tag{6.4}$$

Without loss of generality, (6.2) can be regarded as being the first structural equation, so that $(1 \ \underset{\sim}{\beta}')'$ and $\underset{\sim}{\gamma}$ are the first column of B and Γ respectively. The classical theory of identification assumes that Π in the reduced form (6.3) can always be estimated consistently and is concerned about the nature of restrictions in (6.4) to enable a unique determination of B and Γ given values of Π.

Premultiply (6.2) by the transpose of X to give

$$X'\underset{\sim}{y} = X'Z_1\underset{\sim}{\delta} + X'\underset{\sim}{\mu}\sigma \tag{6.5}$$

then $X'u\sigma$ has mean zero and covariance matrix $\sigma^2X'X$. The conventional two stage least squares (2SLS) estimator of $\underset{\sim}{\delta}$, denoted by d_{2SLS}, is obtained from (6.5) by applying Aitken's theorem. Namely, d_{2SLS} is the unique solution of the normal equation

$$Z_1'X(X'X)^{-1}X'Z_1d_{2SLS} = Z_1'X(X'X)^{-1}X'y \ . \tag{6.6}$$

As X'X is a $K\times K$ square matrix, it will be singular unless the rank of X is K. This requires, among other things, $T > K$. Also, d_{2SLS} exists only if the matrix $Z_1'X(X'X)^{-1}X'Z_1$ has rank n, which requires $K > n$, provided that X has full column rank (i.e. rank K). In econometrics, many large models (e.g. the Brookings Model) have the number of predetermined variables (K) exceeding the number of observations (T) and the rank of X is thus less than K. This is a

situation commonly known as "undersized sample" in the literature. Thus in addition to the usual rank condition ($K \geq n$) for the 2SLS to exist, we now require the extra condition ($K \leq T$). If these latter condition is not fulfilled, modification of the conventional 2SLS is needed because the usual rank condition will be violated as well. Note that this problem is not only confined to 2SLS estimation. Limited-information maximum likelihood, three stage least squares, full-information maximum likelihood and linearized maximum likelihood all directly or indirectly rely on the existence of 2SLS estimates and are hence inapplicable when 2SLS is not applicable.

The problem has been extensively discussed in the literature but the term "undersized" is not always precisely defined. Apart from the above interpretation for the 2SLS case, it refers more loosely to a situation where the number of degrees of freedom in carrying out the first stage of a two stage estimation procedure is very small leading to problems in solving a system of linear equations. A characterization of the undersized sample situation for the FIML estimator is given by Sargan (1975).

For the 2SLS case various modified estimation methods have been suggested of which two are particularly noteworthy. The first involves a *modified* 2SLS (or an instrumental variable procedure) in which the list of predetermined variable is truncated at the initial stage; a second method (2SPC) involves using the principal components of the predetermined variables in 2SLS estimation.[1] Both these types of procedures may be thought of as limiting cases of a class of instrumental variable

---

[1] Truncation methods have been discussed by Fisher (1965a, 1965b) and Mitchell and Fisher (1970). Principal components methods have been discussed by Kloek and Mennes (1960), Amemiya (1966) and Dhrymes (1970). Swamy and Holmes (1971)'s Generalised Inverse technique is trivial extension of 2SPC.

procedures involving continuous shrinkage of the parameter space in the first stage of estimation (see Chapter 2 of this thesis). The method of Generalised Ridge Regression has Restricted Least Squares and the Truncated Principle Components as special cases and can be conveniently thought of as another possible alternative estimator for the first stage of reduced form estimation. Section 2 investigates the properties of a modified 2SLS estimator in which the first stage is carried out by the Generalised Ridge technique. As the motivation to try this approach will exist in small samples only we are especially interested in the derivation of the small sample properties and have used the small-σ asymptotic approach to do so.[2] In Section 3, we motivate the extension of ridge regression technique at both stages of the two stages least squares procedure. A simulation study of the gain for so doing is reported in Section 4 which is followed by a conclusion.

§6.2 Consequence of Using Alternatives to OLS in Stage 1[3]

The results in this section relate to an estimator called Generalised Ridge Instrumental Variable estimator (GRIV). This method is very simply one in which the first stage OLS estimation is replaced by Generalised Ridge method. The relation of this estimator to a number of discrete and continuous shrunken alternatives has been discussed extensively in previous chapters (note especially Chapter 3). We aim here to find out whether the mean square error gains at the first stage of using these alternatives to OLS would carry over to the second stage of estimation. That is, whether the structural estimator of δ displays

---

[2] See Kadane (1971). In view of Sargan's treatment of undersized sample situation, see Sargan (1975) especially Section 2, it would seem that the small σ-asymptotic approach may be especially appropriate in the present context.

[3] Some results of this section were reported in Lee and Trivedi (1979).

smaller mean square error than the conventional 2SLS also. Throughout we shall be concerned with the $T > K$ case. We note that of the four alternatives to OLS in the first stage that we consider, two can be applied when $K > T$, viz. Generalised Ridge and Principal Components. So the conclusions we reach for these cases will also apply in the undersized sample situation. Two other alternatives, viz. the preliminary test estimator and the Stein-James estimator, have no meaningful counterparts in the undersized sample situation. Therefore, our results pertaining to these apply to the undersized samples only in the loose sense of that term.

We consider instrumental variable estimation of equation (6.2) using as instruments the $(T \times n)$ matrix

$$\tilde{Z}_1 = (\tilde{Y}_1, X_1) = (NY_1, X_1) = NZ_1 + \Omega \tag{6.7}$$

where

$$\Omega = \tilde{Z} - NZ_1$$

$$N = X(X'X + HK^*H')^{-1}X'$$

$$K^* = \text{diag}\{k_1^* \ \ldots \ k_n^*\} \ .$$

It is well known that for non-singular A and B, we can write $(A+EBE')^{-1} = A^{-1} - A^{-1}E[B^{-1} + E'A^{-1}E]^{-1}E'A^{-1}$ and let $H'X'XH = \Lambda = \text{diag}\{\lambda_1 \ \ldots \ \lambda_n\}$, we have

$$(X'X+HK^*H')^{-1} = (X'X)^{-1} - (X'X)^{-1}H[K^{*-1} + \Lambda^{-1}]^{-1}H'(X'X)^{-1} \ . \tag{6.8}$$

Thus, it is easy to verify that, for $N_o = X(X'X)^{-1}X'$,

$$N = N_o - X(X'X)^{-1}H[K^{*-1} + \Lambda^{-1}]^{-1}H'(X'X)^{-1}X' \tag{6.9}$$

and

$$N^2 = N - X(X'X)^{-1}H\Lambda^2(\Lambda+K^*)^{-2}K^*H'(X'X)^{-1}X' \ . \tag{6.10}$$

Finally, the GRIV is given by

$$d_{GRIV} = (\tilde{Z}_1'Z_1)^{-1}\tilde{Z}_1'y . \tag{6.11}$$

Now, define $\bar{V}_z = [V_1 : 0]$, $\bar{Z}_1 = [\bar{Y}_1, X_1]$

where

$$Y_1 = \bar{Y}_1 + V_1 \quad \text{and} \quad Z_1 = \bar{Z}_1 + \bar{V}_z \ , \ Q^{-1} = (N\bar{Z}_1 + \Omega)'\bar{Z}_1 = R'\bar{Z}_1 \ ,$$

and $q = \frac{1}{T}\,\&(\bar{V}_z'u)$, we have proved the following results in the Appendix to this chapter.

Theorem 6.2.1.

The bias of d_{GRIV} is given by

$$\&(d-\delta) = \sigma^2[(\text{tr } N)Q - Q\bar{Z}_1'NRQ - (\text{tr } N\bar{Z}_1QR')Q]q + O(\sigma^3) \ . \tag{6.12}$$

This result may be viewed as an extension of one of the results contained in Kadane's paper. Essentially it makes the bias dependent upon the ridge parameters in the matrix K^* through the matrix N.

Theorem 6.2.2.

If the ridge parameters are chosen such that $K^* = \sigma^2\bar{K}$ where $\bar{K}$ is a constant diagonal matrix independent of σ^2, then we can write equation (6.10)

$$N^2 = N + \sigma^2 N_1$$

where

$$N_1 = -X(X'X)^{-1}H\Lambda^2(\Lambda+\sigma^2\bar{K})^{-2}\bar{K}H'(X'X)^{-1}X'$$

and the mean square error d_{GRIV} is given, to $O(\sigma^4)$, by

$$\&(d-\delta)(d-\delta)' = \sigma^2Q + \sigma^4Q(\bar{Z}_1'N_1\bar{Z}_1 + A^*)Q + O(\sigma^5) \tag{6.13}$$

where

$$A^* = (\text{tr } QC_1)Q^{-1}(2n+3-2 \text{ tr } N) + (\text{tr } C_2Q)Q^{-1}$$
$$+ [(n+2-\text{tr } N)^2 + 2(\text{tr } N^2) - 2 \text{ tr } N + 2]C_1 \qquad (6.14)$$
$$+ [\text{tr } N^2 - 2 \text{ tr } N + n + 2]C_2 \ .$$

(C_1 and C_2 are defined in the Appendix).

The condition $K^* = 0(\sigma^2)$ has been found mathematically convenient in our derivation but its motivation can be traced back to some good selection criteria for the ridge parameters. As we know, the M.S.E. - optimal choice for k_i^* is $\frac{\sigma^2}{\gamma_i^2}$ which is clearly $0(\sigma^2)$ (where γ_i's are the unknown true principal components of the regression coefficients in stage 1). Other selection criteria such as Allen's PRESS and Mallow's C_p, both make K^* a stochastic magnitude and suggest that $K^* = 0(\sigma^2)$ is a practically relevant choice.[4]

In order to make comparisons between different estimators we note that the differences arise from differences in the choice of N and the resultant matrix Q. They are listed below in Table 6.1.1 for reference.

---

[4] If, however, K^* is chosen in a way which renders it stochastic, then the expressions for bias and mean square error given in this section will not apply. Also this choice of K^* would preserve consistency and asymptotic efficiency because $\sigma^2 \to 0$ as $T \to 0$, hence GRIV is asymptotically equivalent to 2SLS.

Table 6.1.1

Differences in N and Q for Various Alternatives to 2SLS

Estimator	N	tr N	Q^{-1}
Classical			
OLS	I_T	T	$\bar{Z}_1'\bar{Z}_1$
2SLS	N_o	K	$\bar{Z}_1'\bar{Z}_1$
k-class	$(1-k)I_T+kN_o$	$(1-k)T+kK$	$\bar{Z}_1'\bar{Z}_1$
Alternatives			
2SPC(h)	$S(S'S)^{-1}S'$	$h+K_1$	$\bar{Z}_1'S(S'S)^{-1}S'\bar{Z}_1$
GRIV	$X(X'X+HK^*H')^{-1}X'$	$\sum_{i=1}^{K} \frac{\lambda_i}{\lambda_i+k_i^*}$	$(N\bar{Z}_1+\Omega)'\bar{Z}_1$
James-Stein	$(1-h(\cdot))N_o$	$(1-h(\cdot))K$	$(1-h(\cdot))\bar{Z}_1'\bar{Z}_1$
PTE	$N_o+I_{(0,c^*)}(w)[S(S'S)^{-1}S'-N_o]$	$K+I_{(0,c^*)}(w)[h-K_2]$	$\bar{Z}_1'N\bar{Z}_1$

Please see below and the Appendix for definitions on notation.

2SPC(h) stands for the Kloek-Mennes Two-Stage Principal Components estimator (X has assigned rank $h \leq K$) so that $S = [XH_h, X_1]$ where H_h, a $(K\times h)$ matrix, consists only of the first h columns of the $(K\times K)$ matrix H.

Note also that $\Omega = (0,(I-N)X_1)$ can be considered as the residual of X_1 after a regression by generalised ridge method into the space of X. This term is significant if we insist on using X_1 rather than NX_1 as an instrumental variable for X_1, but which will be identically zero for 2SLS and 2SPC because in these cases $I-X(X'X)^{-1}X'$ and $I-S(S'S)^{-1}S'$ are idempotent and orthogonal to X_1. A similar property can be imposed on the GRIV if we can find a $(T\times h)$ matrix F such

that $N = X(X'X+HK^*H')^{-1}X' \equiv \tilde{S}(\tilde{S}'\tilde{S})^{-1}\tilde{S}'$ where $S = (F : X_1)$ so that $I - N$ is again orthogonal to X_1. In all such cases, $\Omega \equiv 0$, $R \equiv N\bar{Z}_1$ and $N^2 \equiv N$.

(A) Bias Comparison

Comparison of the bias of 2SLS and 2SPC is straightforward. They are given respectively by $\sigma^2[K-1-n](\bar{Z}_1'\bar{Z}_1)^{-1}q$ and $\sigma^2[(h+K_1)-1-n](\bar{Z}_1'S(S'S)^{-1}S'\bar{Z}_1)^{-1}q$ since $Q\bar{Z}_1'NR \equiv I_n$ in both cases. Though $h - G_1 - 1$ is less than $K_2 - G_1 - 1$ we note that

$$(\bar{Z}_1'S(S'S)^{-1}S'\bar{Z}_1)^{-1} - (\bar{Z}_1'\bar{Z}_1)^{-1} \tag{6.15}$$

is positive semidefinite so that a small bias though achievable, is not guaranteed for 2SPC. This result casts doubt on Amemiya (1966)'s optimism about 2SPC.

A conclusion similar to the one above holds for the bias of GRIV estimator when $N \equiv N+\sigma^2N_1$. In that case, the bias up to $O(\sigma^2)$ is $[(\text{tr } N) - 1 - (\text{tr } I_n)](\bar{Z}_1'N\bar{Z}_1)^{-1}q$. Thus, when compared with the bias of 2SLS, it is possible that

$$\sum_{i=1}^{K} \frac{\lambda_i}{\lambda_i+k_i^*} - (K_1 + G_1 + 1) < K_2 - G_1 - 1 \tag{6.16}$$

but

$$(\bar{Z}_1'X(X'X + HK^*H')^{-1}X'\bar{Z}_1)^{-1} - (\bar{Z}_1'\bar{Z}_1)^{-1} \tag{6.17}$$

is positive definite.

For the James-Stein estimator, we have

$N = [1-h(\cdot)]N_o$ *where* $0 < h(\cdot) < 1$ is a real value function of $O(\sigma^2)$

$= N_o + O(\sigma^2)$ $\qquad N_o = X(X'X)^{-1}X'$.

Exact comparison with 2SLS are possible only when $h(\cdot)$ is non-stochastic as in the case of Mayer and Willke (1973)'s shrunken estimator. In the latter case the bias of this variant will be up to $O(\sigma^2)$, the same as that of 2SLS if $h(\cdot)$ is of $O(\sigma^2)$.

For the preliminary test estimator (PTE), the choice between 2SLS and the restricted 2SLS (such as that of truncation of small principal components) depends on whether a test statistic w falls in a pre-specified critical region. That is, N is a stochastic function of w so that

$$N = N_o + I_{(0,c^*)}(w)[S(S'S)^{-1}S' - N_o] \tag{6.18}$$

where $I_{(0,c^*)}(w)$ is the indicator function which is 1 if w falls in the region indicated.

For $0 < c^* < \infty$, there is a possibility of bias reduction over 2SLS, as in the 2SPC case.

(B) MSE Comparison

Specialising Theorem 2 for different estimators we obtain expression for MSE of d. Comparison may be based on the $O(\sigma^2)$ term alone if that term differs between estimators. In the event that the term of $O(\sigma^2)$ is the same we may consider the term of $O(\sigma^4)$. Comparing 2SPC with 2SLS it is readily seen from terms of $O(\sigma^2)$ that the MSE of the former exceeds that of the latter since the difference between the two Q matrices is

$$(\bar{Z}_1'S(S'S)^{-1}S'\bar{Z}_1)^{-1} - (\bar{Z}_1'\bar{Z}_1)^{-1}$$

which is positive semidefinite. Similarly, MSE of GRIV exceeds that of 2SLS since

$$\bar{Z}_1'\bar{Z}_1 - \bar{Z}_1'N\bar{Z}_1 = \bar{Z}_1'[X(X'X)^{-1}H(\Lambda^{-1} + K^{*-1})^{-1}H'(X'X)^{-1}X']\bar{Z}_1 \qquad (6.19)$$

is positive semidefinite.

Comparison between any pair of the triple (2SLS, James-Stein, PTE) is difficult because the latter two imply a stochastic K^* for which our MSE expression is not strictly valid. Provided however that we can assume $K^* = 0(\sigma^2)$, James-Stein estimator and 2SLS will have the same MSE to $0(\sigma^2)$. In the light of previous comparisons we also conjecture that MSE of the PTE variant will exceed that of 2SLS, but would be less than that of 2SPC. Recall that for the reduced form estimation, neither PTE nor PCE has uniformly lower risk over the entire parameter space, we have here more definite to say about the relation among the estimator of the structural coefficients.

Finally we note that no ranking between GRIV and 2SPC can be established since the final result depends very much on the particular choice of k_i^*. Of course this ambiguity cannot be resolved by considering terms of $0(\sigma^4)$.

In summary, estimators that are known to have *gains* in MSE over OLS in the first stage may turn out to have *losses* in MSE over 2SLS in the second stage. Attempts to improve on 2SPC by GRIV techniques, when 2SLS is not feasible, could quite possibly produce estimates with larger mean square error depending upon the choice of K^*. The results can be extended to the Theil's (1973) D-class estimator which is a crude modification of the 2SLS, using

$$N = XD^{-1}X' \qquad (6.20)$$

where D = a diagonal matrix whose diagonal elements are the same as the diagonal elements of $X'X$.

It is worth considering the possibility that application of

shrinkage technique at the first and the second stage of estimation could be a source of mean square error gains. This is the topic we study next.

§6.3 "Improved" Two Stage Least Squares

Although there is a hierarchy of system methods for estimating the structural coefficients of a simultaneous equation model ranging from the highly nonlinear solution of a maximised likelihood function (with all a priori information incorporated) to the naive application of OLS ignoring all other restrictions and/or serial correlated disturbances, 2SLS is by far the most popular technique among economists. As is well known, 2SLS is simple computationally, and is consistent and asymptotically efficient in the class of regular estimators. The theoretical rankings of estimators based on asymptotic theory are generally ignored in practice because of the usually small sample sizes of data and, more importantly, because of the general poor performance of the more expensive methods. One explanation is that the ideal ranking assumes that the model to be estimated is perfectly specified which is not true in practice. Even the "limited information" methods 2SLS and 2SPC are affected by specification errors in other equations if they utilize all the exogenous variables in the model for obtaining the proxies in the first stage. It is not surprising that some econometricians are prepared to suffer a little inconsistency and use OLS for structural estimation. If an estimator, though theoretically inconsistent, produces more plausible economic results in simulations or in real data situations, it would be good enough justification for their use.

Multicollinearity in the data is possibly the most commonly neglected factor in considering the choice of various estimation techniques. Klein and Nakamura (1962) give a systematic analysis of the sources and consequence of collinearity in system methods. Their results

can be conveniently expounded in terms of the general k-class estimator for the structural coefficients in equation (6.2). Let us define S_1 as the selection matrix of order $(G+K)\times(G+K)$ such that

$$Z_1 \equiv ZS_1 = [Y_1, X_1] \tag{6.21}$$

and the $(G+K)\times 1$ selection vector s_1 to pick up the particular dependent variable from the full data matrix

$$\underset{\sim}{y} = Zs_1 \tag{6.22}$$

where

$$Z = (Y, X) .$$

Thus, we obtain the k-class estimator as the solution d_k of the system of linear equations

$$S_1'[(1-k)Z'Z + kR]S_1 d_k = S_1'[(1-k)Z'Z + kR]s_1 \tag{6.23}$$

where

$$R = Z'X(X'X)^{-1}X'Z$$

k is a scalar constant .

It is well known that when $k = 1$, d_1 is the 2SLS, when $k = 0$, d_o is the OLS. Let $\hat{\lambda}$ be the smallest root of $\det|Y_o'M_1Y_o - \lambda Y_o'MY_o| = 0$ *where*

$$Y_o = [\underset{\sim}{y}, Y_1], \quad M_1 = I - X_1(X_1'X_1)^{-1}X_1' \quad \text{and} \quad M = I - X(X'X)^{-1}X' . \tag{6.24}$$

The Limited-Information maximum likelihood estimator (LIML) has $k = \hat{\lambda}$.

In general, d_k exists and is unique if the matrix $S_1'[(1-k)Z'Z + kR]S_1$ is non-singular. However, in general, non-experimentally generated economic data would be collinear and would cause this matrix to be ill-conditioned. It is not hard to explain the observation that 2SLS are more sensitive to the presence of

multicollinearity than are OLS and that LIML are more sensitive to multicollinearity than are 2SLS.

(i) When $k = 1$, the matrix to be inverted at (6.23) becomes

$$S_1'RS_1 = S_1'Z'X(X'X)^{-1}X'ZS_1$$

$$= [\hat{Y}_1, X_1]'[\hat{Y}_1, X_1] \qquad (6.25)$$

where

$$\hat{Y}_1 = X(X'X)^{-1}X'Y_1 \ . \qquad (6.26)$$

If $K < T$, then $\hat{Y}_1$, being a projection of Y_1 to the K-dimensional space of X, must have a rank smaller than Y_1. Thus, the rank of $S_1'RS_1$ must be smaller than or equal to that of $S_1'Z'ZS_1$. If the latter matrix is ill-conditioned, the former must also be ill-conditioned but not vice versa. Thus, 2SLS is relatively more unstable than OLS in general.

(ii) If $K > T$, then the problem of undersized sample is said to exist. 2SLS in this situation is identical to direct OLS. (See Swamy and Holmes (1971)).

(iii) The 2SLS estimator for β $(k = 1)$ can be written as

$$[\hat{Y}_1, X_1]'[\hat{Y}_1, X_1] \begin{pmatrix} b \\ c \end{pmatrix} = [\hat{Y}_1, X_1]'y_1 \qquad (6.27)$$

or

$$\hat{Y}_1'M_1\hat{Y}_1 b = \hat{Y}_1 M_1 y \ . \qquad (6.28)$$

But, it is easy to verify that the matrix to be inverted is

$$\hat{Y}_1'M_1\hat{Y}_1 = Y_1'(M_1 - M)Y_1 \qquad (6.29)$$

where

M_1 and M are defined at (6.24).

If X_2 adds very little to the explanation of the variation of Y_1 after the effect of X_1 has been removed, then M_1Y_1 will be very close to MY_1 causing the matrix (6.29) to be ill-conditioned. This happens even if X_1 and X_2 are orthogonal, providing an additional source of multicollinearity for 2SLS. It is therefore important that the original system must be specified in such a way that it is "strongly" identified, i.e. the predetermined variables not appearing in the equation must be responsible for the variation of those endogenous variables included as regressors in the equation in order that 2SLS be well determined.

(iv) In the case of LIML, we can write $\lambda = 1+\mu$ so that μ is the smallest root of the determinantal equation

$$\det\left|Y_o'(M-M_1)Y_o - \mu Y_o'MY_o\right| = 0 \ . \qquad (6.30)$$

It is possible that $Y_o'(M-M_1)Y_o$ is singular and the smallest μ will be zero so that LIML is identical to 2SLS in this case. If that matrix is non-singular, but the smallest positive root is multiple, then there exist two linear independent e-vectors corresponding to the smallest e-root, resulting in ambiguity in choosing the estimates. Even if we only have a simple smallest e-root, $(Y_1'(M-M_1)Y_1 - \mu Y_1'MY_1)$ may happen to be singular, then the solution of the homogeneous equation

$$(Y_1'(M-M_1)Y_1 - \mu Y_1'MY_1)b = 0 \qquad (6.31)$$

would be indeterminate.

In general, it is more likely to encounter difficulty of inverting singular matrices in LIML than in 2SLS because singularity of the former implies that of the latter, while the converse is not necessarily true.

There are only a few attempts to tackle this problem of collinearity in system methods. If the problem occurs among the predetermined variables, only the first stage estimation need be adjusted and their implications have been analysed in the preceding section. If all the variables exhibit certain degree of collinearity, then remedy at only the first stage is not adequate. Neeleman (1973) was perhaps the first econometrician who tackle the problem directly. He proposed a so-called Generalised Two Stage Least Squares procedure (G2SLS) which entails repeated use of generalised inverse technique for the inversion of singular matrices in both stages. However, in practice, the matrices to be inverted are not exactly singular but nearly so, G2SLS is therefore identical to 2SLS. In order to avoid this, Neeleman proposed replacing the near-singular matrix by an approximate matrix of less than full rank. The approximation is done by choosing a linear combination of a subset of the columns of the original matrix such that the norm of the difference between the new matrix and the original one is minimum. This turns out to be an expensive way of getting truncated principal components estimate at both stages. An alternative solution for ill-condition data is to incorporate a priori information. Chipman (1964) obtained the Minimum Average Risk Linear (MARL) estimator for the coefficients of the classical multiple regression model when there is prior information about the location and dispersion of the coefficients. Consider the linear model

$$\underset{(T\times 1)}{y} = \underset{(T\times p)}{X}\,\underset{(p\times 1)}{\beta} + \underset{(T\times 1)}{u} \qquad \mathcal{E}u = 0 \qquad \mathcal{E}uu' = \sigma^2\Omega \tag{6.32}$$

with the prior information $\beta - \bar{\beta} = w$ such that

$$\&w = 0 \quad \text{and} \quad \text{var}(w) = \tau^2\Theta \ . \tag{6.33}$$

Note that we have slightly altered the conventional treatment of prior information by treating the prior mean $\bar{\beta}$ as random. Then, Chipman's MARL is equivalent to the Theil-Goldberger's mixed estimator for β obtained by applying Aitken's procedure to the following model

$$\begin{pmatrix} y \\ \bar{\beta} \end{pmatrix} = \begin{pmatrix} X \\ I \end{pmatrix} \beta + \begin{pmatrix} u \\ w \end{pmatrix}, \ \text{var} \begin{pmatrix} u \\ w \end{pmatrix} = \begin{pmatrix} \sigma^2\Omega & 0 \\ 0 & \tau^2\Theta \end{pmatrix} \tag{6.34}$$

giving

$$\tilde{\beta}_{MARL} = \left(\frac{1}{\sigma^2} X'\Omega^{-1}X + \frac{1}{\tau^2} \Theta^{-1} \right)^{-1} \left(\frac{1}{\sigma^2} X' \ ^{-1}y + \frac{1}{\tau^2} \Theta^{-1}\bar{\beta} \right) . \tag{6.35}$$

If Θ is singular, then using the identity

$$\frac{1}{\sigma^2} X'\Omega^{-1}(\tau^2 X\Theta X' + \sigma^2\Omega) \equiv \left(\frac{X'\Omega^{-1}X}{\sigma^2} + \frac{1}{\tau^2}\Theta^{-1} \right) \tau^2\Theta X' \tag{6.36}$$

one can obtain

$$\begin{aligned} \tilde{\beta}_{MARL} &= \tau^2\Theta X'(\tau^2 X\Theta X' + \sigma^2\Omega)^{-1}y + [I - \tau^2\Theta X'(\tau^2 X\Theta X' + \sigma^2\Omega)^{-1}X]\bar{\beta} \\ &= \Theta X'(X\Theta X' + k\Omega)^{-1}y + [I - \Theta X'(X\Theta X' + k\Omega)^{-1}X]\bar{\beta} \end{aligned} \tag{6.37}$$

where $k = \dfrac{\sigma^2}{\tau^2}$.

In econometrics, σ^2 is expected to be finite and positive, so that $k \to 0$ if $\tau^2 \to \infty$. One can see from (6.35) that the limiting MARL estimator approaches the Direct Generalised Least Square estimator when the prior variance of the regression coefficients becomes infinite. Note that the Straight ridge estimator has $\Theta \equiv I_p$ and $\bar{\beta} = 0$ whereas the Shiller estimator has $\Theta = R'R$ and $\bar{\beta} = 0$. Furthermore, consistency

of the estimator $\tilde{\beta}_{MARL}$ requires that $(X'\Omega^{-1}X)^{-1}$ exist and approach the zero matrix as T approaches infinity, but $(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$ approaches β. In this case, we can expand the left hand matrix as a first order Taylor Series,

$$\begin{aligned}(X'\Omega^{-1}X+k\Theta^{-1})^{-1}X'\Omega^{-1}y &= [I+k(X'\Omega^{-1}X)^{-1}\Theta^{-1}]^{-1}(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y \\ &\doteq \{I-k(X\Omega^{-1}X)^{-1}\Theta^{-1}\}(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y \\ &\doteq (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y \quad \text{as} \quad k(X'\Omega^{-1}X)^{-1}\Theta^{-1} \to 0 .\end{aligned}$$

Thus, if k is independent of the sample size T, the estimator $\tilde{\beta}_{MARL}$ is asymptotically equivalent to the Generalised Least Square Estimator $\hat{\beta}_{GLS}$. This is a particularly useful result as it is clear that the asymptotic properties of consistent estimators for structural equations are preserved by ridge-method.

Swamy and Mehta (1976) applied Chipman's MARL estimator to estimating the structural parameters δ of the transformed equation (6.5). Since the disturbance term has a covariance matrix equal to $\sigma^2X'X$, then when one has prior information $\delta - \bar{\delta} = w$ with $\mathcal{E}w = 0$ and $\text{var}(w) = \Delta$, the MARL-like estimator for δ is

$$d_{SM1}\begin{cases} = \left(\dfrac{Z_1'X(X'X)^{-1}X'Z_1}{\sigma^2}+\Delta^{-1}\right)^{-1}\left(\dfrac{Z_1'X_1(X'X)^{-1}X'y_1}{\sigma^2}+\Delta^{-1}\bar{\delta}\right) & (6.38) \\ \qquad \text{if } \Delta \text{ is nonsingular} & \\ = \Delta Z_1'X(X'Z_1\Delta Z_1'X+\sigma^2X'X)^{-1}y_1+[I-\Delta Z_1'X(X'Z_1\Delta Z_1'X+\sigma^2X'X)^{-1}Z_1X']\bar{\delta} & (6.39) \\ \qquad \text{if } \Delta \text{ is singular .} & \end{cases}$$

The above estimator is not feasible when the sample is under-sized, where typically $X'X$ will be singular or nearly so. In the circumstance, Swamy and Mehta suggested using the generalised-inverses

in place of the proper inverses that appear in expression (6.39), namely

$$d_{SM2} = \bar{\delta} + \Delta Z_1' X[X(Z_1 \Delta Z_1 + \sigma^2 I)X]^{-} X'(y_1 - z_1 \bar{\delta}) \ . \qquad (6.40)$$

They claimed that $X[X'(Z_1 \Delta Z_1 + \sigma^2 I)X]^{-} X'$ is invariant for any generalised inverse of $[X'(Z_1 \Delta Z_1 + \sigma^2 I)X]$ and that the estimator (6.40) exists even when Δ is singular and Z_1 has less than full column rank. Since Chipman's theorem is not fully applicable to an equation with stochastic regressors, one must rely on the result of Zellner and Vandaele (1974) that if both the sample and prior distributions are normal, the estimator d_{SM1} is approximately equal to the mean of the conditional posterior distribution. The estimator also possesses finite second order moments regardless of the degree of over-identification.

If one simply use $\Delta = \frac{k}{\sigma^2} I$ and $d = 0$, then the formula for d_{SM1} at (6.38) is applicable but with $(X'X)^{-1}$ replaced by the generalised inverse matrix to give the ridge-type two stage least squares

$$d_{SM3} = (Z_1' X(X'X)^{-} X' Z_1 + kI)^{-1} Z_1' X(X'X)^{-} X' y \qquad k \geq 0 \qquad (6.41)$$

when k becomes zero as information is diffuse, d_{SM3} tends to the Swamy and Holmes (1971) estimator. We have pointed out in the last section that the latter is not formally different from the 2SPC if the assigned rank of X is $h < K$ and from OLS if $K \geq T$. One can in fact apply the ridge-type technique also in the first stage to cope with undersized sample problem and to obtain an estimator that is asymptotically equivalent to the conventional 2SLS. Such approach is sensible because the first stage predictors from the reduced form could have been improved upon by a ridge regression anyway, which are then used as instrumental variables for the second stage. The application of the ridge regression is thus justified at the second stage as a means to

improve point estimates for individual structural coefficients since least squares procedures are impaired by possible multicollinearity. Neeleman (1973) applied the generalised-inverse technique at both stages and his Monte Carlo studies indicate that substantial improvement in mean square error is possible by his method. It would be instructive to see what combination of modifications to 2SLS would best reduce the mean square error. We have therefore conducted a Monte Carlo study to compare the performance of nine possible variants of the two stage procedures in a small sample situation, including 2SLS and 2SPC as special cases.

§6.4 Finite Sample Properties

Mariano (1972) has shown that when certain conditions are satisfied, the even moments of the conventional 2SLS of $\underset{\sim}{\delta}$ exist only up to an order not exceeding the degree of over-identification of the particular structural equation to be estimated. Sawa (1972) has derived explicit formulae for the mean and variance of the conventional 2SLS as a member in the k-class estimator. The results are obtained for an equation which is over-identified by degrees of at least 2, in a system that consists of only two endogenous variables. Mehta and Swamy (1978) established that their ridge-type two stage least squares estimator d_{SM3} at (6.41), for $k > 0$, possesses the first two moments regardless of the degree of over-identification. For a model with two endogenous variables, the formulae for mean and variance of d_{SM3} could be obtained in the same spirit as Sawa. This is one major justification for using d_{SM3} as an alternative to d_{2SLS} as the latter does not, in general, possess finite second order moments and has unbounded risk under a

quadratic loss function if the degree of over-identification of the equation is less than or equal to unity. We now extend these results to our modified two stage least square with ridge regression at *both* stages.

Let k_1 and k_2 be the ridge parameters for the first and second stage respectively. Then the $RR(k_2)RR(k_1)$ estimator is the solution of the following linear equations

$$\begin{pmatrix} Y_1'X(X'X+k_1I)^{-1}X'Y_1+k_2I & Y_1'X_1 \\ X_1'Y_1 & X_1'X_1+k_2I \end{pmatrix} \begin{pmatrix} b \\ c \end{pmatrix} = \begin{pmatrix} Y_1'X(X'X+k_1I)^{-1}X'y \\ X_1'y \end{pmatrix} . \quad (6.42)$$

Expanding along the second row, we have

$$c = (X_1'X_1 + k_2I)^{-1}X_1'(y-Y_1b) . \quad (6.43)$$

Substituting back to the first row, we obtain

$$b = [Y_1'\tilde{M}Y_1 + k_2I]^{-1}Y_1'\tilde{M}y \quad (6.44)$$

where

$$\tilde{M} = X(X'X + k_1I)^{-1}X' - X_1(X_1'X_1 + k_2I)^{-1}X_1' . \quad (6.45)$$

The validity of Mehta and Swamy's result hinges on the matrix $\tilde{M}$ being positive semidefinite of rank K and hence has positive e-roots. We now establish,

Theorem 6.4.1.

The matrix $\tilde{M}$ is positive semidefinite of rank K and has K positive e-roots provided $k_2 > k_1 > 0$.

Proof.

$$\tilde{M} = X(X'X+k_1I)^{-1}X' - X_1(X_1'X_1+k_2I)^{-1}X_1'$$

$$= (I-X_1(X_1'X_1+k_2I)^{-1}X_1') - (I-X(X'X+k_1I)^{-1}X') \ .$$

Let $D = \begin{pmatrix} I \\ 0 \end{pmatrix}(I \quad 0)$ and $X_1 = X \begin{pmatrix} I \\ 0 \end{pmatrix}$, then

$$I-X_1(X_1'X_1+k_2I)^{-1}X_1' = k_2(X_1'X_1+k_2I)^{-1}$$

$$= k_2[XDX'+k_2I]^{-1}$$

$$= I-X(X'X)^{-1}X' + k_2X(X'X)^{-1}[k_2(X'X)^{-1}+D]^{-1}(X'X)^{-1}X'.$$

On the other hand,

$$I-X(X'X+k_1I)^{-1}X' = I-X\{(X'X)^{-1} - (X'X)^{-1}(\frac{1}{k_1} I + (X'X)^{-1})^{-1}(X'X)^{-1}\}X'$$

$$= I-X(X'X)^{-1}X' + k_1X(X'X)^{-1}[k_1(X'X)^{-1}+I]^{-1}(X'X)^{-1}X'.$$

Thus,

$$\tilde{M} = k_2X(X'X)^{-1}[k_2(X'X)^{-1}+D]^{-1}(X'X)^{-1}X'$$

$$- k_1X(X'X)^{-1}[k_1(X'X)^{-1}+I]^{-1}(X'X)^{-1}X'$$

$$= X(X'X)^{-1}\{k_2[k_2(X'X)^{-1}+D]^{-1} - k_1[k_1(X'X)^{-1}+I]^{-1}\}(X'X)^{-1}X'$$

$$= X(X'X)^{-1}[k_2(X'X)^{-1}+D]^{-1}\{k_2[k_1(X'X)^{-1}+I] - k_1[k_2(X'X)^{-1}+D]\}[k_1(X'X)^{-1}+I]^{-1}$$
$$(X'X)^{-1}X'$$

$$= X(X'X)^{-1}[k_2(X'X)^{-1}+D]^{-1}\begin{Bmatrix} (k_2-k_1)I_{K_1} & 0 \\ 0 & k_2I_{K_2} \end{Bmatrix}[k_1(X'X)^{-1}+I]^{-1}(X'X)^{-1}X' \ .$$

As long as $k_2 > k_1 > 0$, $\tilde{M}$ is seen to be a positive semidefinite matrix of rank K.

Q.E.D.

Exact finite sample moment formula for the various modified versions of 2SLS are difficult to obtain and even if we succeed in getting them out, they are too complicated to throw any light on the problem of discriminating among competitors. An alternative to this labour - intensive analytic comparison is the relatively simple approach through computer simulation. We present here results of a small Monte Carlo study to assess the performance of nine variants of the modified 2SLS estimator for δ in a particular structural equation, since there are three different possible alternatives that could be used in each stage, namely, Least Squares (LS), Principal Components (PC) and Ridge Regression (RR). It is certainly true that our results are specific to the model size and structure we have used. However, the results provide reasonably meaningful indications about the relative merits of the estimators and would serve as a pilot study for arriving at a tentative conclusion.

To facilitate comparison with previous results in the literature, the design of the simulation model follows closely that of Neeleman (1973) which was based on Mosbaek and Wold's (1970) two equations model. Symbolically, the model is

$$\begin{pmatrix} 1 & \beta_{12} \\ \beta_{21} & 1 \end{pmatrix} \begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} + \begin{pmatrix} \gamma_{11} & \gamma_{12} & 0 & 0 \\ 0 & 0 & \gamma_{23} & \gamma_{24} \end{pmatrix} \begin{pmatrix} x_{1t} \\ x_{2t} \\ x_{3t} \\ x_{4t} \end{pmatrix} = \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix} \qquad (6.46)$$

$$t = 1 \ldots T$$

It is easy to check that both equations are over-identified as is common in econometrics.

The exogenous variables are sampled from a multinormal distribution with x_{1t} and x_{2t} strongly correlated, otherwise all variables

are independent with zero mean, and the covariance matrix is

$$\begin{pmatrix} 1 & .9 & 0 & 0 \\ .9 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{6.47}$$

20 observations of the exogenous variables are drawn and is then fixed for the rest of 50 replications to generate the endogenous variables from the reduced form.

The error terms of the reduced form are drawn for each replication from a 2-dimensional normal distribution with zero mean and covariance matrix $\sigma^2 I_2$. Following Neeleman, σ^2 has been set to equal .2.

For simplicity, the coefficients of the endogenous variables are assumed to be equal, but opposite in sign so that $\beta_{12} = -\beta_{21} = \beta$. For the same reason, the coefficients of the exogenous variables are assumed to be equal two by two, so that $\gamma_{11} = \gamma_{12} = \gamma_1$ and $\gamma_{23} = \gamma_{24} = \gamma_2$.

As the variances of the endogenous variables are related to the structural coefficients by, through the reduced forms,

$$\text{var}(y_{1t}) = \left(\frac{1}{1+\beta^2} \right)^2 \text{var}(-\gamma_1 x_{1t} - \gamma_1 x_{2t} + \beta\gamma_2 x_{3t} + \beta\gamma_2 x_{4t}) + \text{var}(v_{1t})$$

$$\text{var}(y_{2t}) = \left(\frac{1}{1+\beta^2} \right)^2 \text{var}(-\beta\gamma_1 x_{1t} - \beta\gamma_1 x_{2t} - \gamma_2 x_{3t} - \gamma_2 x_{4t}) + \text{var}(v_{2t})$$

where v_{1t} and v_{2t} are the reduced form disturbances for the two equations variables.

Restricting $\text{var}(y_{1t}) = \text{var}(y_{2t}) = 1$, we can deduce that

$$\gamma_1 = 2\sqrt{\frac{1+\beta^2}{19}} \quad \text{and} \quad \gamma_2 = 2\sqrt{\frac{1+\beta^2}{10}} \tag{6.48}$$

This normalisation facilitates the study of the influence of the size of β as a source of multicollinearity among the endogenous variable and the exogenous variables. Once β is specified, implied values of γ_1 and γ_2 can be obtained from the relations at (6.48). Our attention is focused on the case $\beta = .5$. For each sample of 20 observations generated, nine estimators involving the use of LS, PC or RR at either of the two stages are computed and their mean square errors are compared. Note that the number of principal components retained in PC is determined by the criterion that the sum of the e-roots corresponding to the retained components accounted for at least 95% of the total sum, while the ridge parameters are chosen according to Obenchain's minimum SSCBC criterion. Of course, there are many other optimality criteria for deciding which components to be truncated and ridge parameter to be selected. These are done only for the first replicate and then fixed for the rest of the experiment. For the first structural equation, the first replicate indicates that $k_1 = 7$ and $k_2 = 11$ are good choices of ridge parameters for the first and second stage respectively. As $k_2 > k_1$ in this case, we know that the estimator with ridge regression at both stages possesses the first and second moments and it makes sense to consider its bias and mean square error in finite sample situations. In addition to the results obtained for the coefficients of the first structural equation, we have also obtained the bias and mean square errors for the coefficients of the second reduced form equation, the latter represents the more familiar situation of multiple regression where theoretical results about alternatives to Least Squares have been studied extensively in previous chapters of this thesis. Results of the experiment are summarised in Tables 6.4.1 and 6.4.2 respectively.

Table 6.4.1

The Mean Square Error and Bias of Nine Different Estimators for the Coefficients of the First Structural Equation

(Sample Size = 20 No. of replicates = 50)

Estimators	2nd stage	1st stage		β_{12} .5	γ_{11} .513	γ_{12} .513	TMSE (Total)
d_1	PC	LS	MSE	.0124	.0125	.0647	.0896
			bias	.0320	.0788	-.2525	
d_2	LS	LS	MSE	.0120*	.0226	.2052	.2398
			bias	.0294+	.0307+	-.1151+	
d_3	RR(11)	LS	MSE	.0527	.0130	.1052	.1709
			bias	-.2200	-.1021	-.3195	
d_4	PC	PC	MSE	.0125	.0116*	.0581	.0822*
			bias	.0400	.0766	-.2387	
d_5	LS	PC	MSE	.0124	.0263	.1871	.2258
			bias	.0360	.0593	-.1911	
d_6	RR(11)	PC	MSE	.0514	.0129	.1078	.1721
			bias	-.2172	.0997	-.3241	
d_7	PC	RR(7)	MSE	.0466	.0121	.0437*	.1024
			bias	.1721	.0781	-.2059	
d_8	LS	RR(7)	MSE	.0457	.0261	.1810	.2528
			bias	.1688	.0652	-.1699	
d_9	RR(11)	RR(7)	MSE	.0497	.0147	.1051	.1695
			bias	-.2125	-.1087	-.3199	

* The minimum MSE within column is asterisked.

+ The minimum bias within column is daggered.

Table 6.4.2

The Mean Square Error and Bias of Three Different Estimators for the Coefficients of the Second Reduced Form Equation

Estimator	True Value	$-\Pi_{21}$ .2052	$-\Pi_{22}$ .2052	$-\Pi_{23}$ .5657	$-\Pi_{24}$ .5657	TMSE (Total)
LS	MSE	.0186	.1545+	.0072+	.0128*	.1931
	bias	-.0134	-.0131+	-.0085+	.0233	
PC	MSE	.0064	.0310	.0068*	.0310+	.0572*
	bias	.0434	-.1742	.0148	.0136+	
RR(7)	MSE	.0030*+	.0268*	.0184	.0174	.0656
	bias	.0030+	-.1460	-.1207	-.1031	

* The minimum MSE within column is asterisked.

\+ The minimum bias within column is daggered.

As the first two moments do not exist in some of the estimators considered, it is argued that a measure of accuracy by the mean absolute error (MAE) is preferred to that by the mean square error (MSE).

Following Summers (1965, p.13), we carry out pairwise comparison of the relative accuracy among the nine different estimators using a non-parametric test of the equality of MAE's. Since the null hypothesis H_o: $MAE(d_i) = MAE(d_j)$ is equivalent to the null hypothesis that H_o^*: $p_{ij} = \frac{1}{2}$ *where*

$$MAE(d_k) = \frac{\sum_{t=1}^{N} |d_{k(t)} - \theta|}{N} \qquad k = 1,\ldots,9 \text{ estimators}$$

$N = 50$ replicates

θ = true value

t = replicate index.

$$P_{ij} = \text{Prob}\{|d_i - \theta| > |d_j - \theta|\} \qquad i,j = 1,\ldots,9$$

Table 6.4.3

Summer's Test Statistic on MAE (Z-scores)

(Pairwise Comparison of Various Estimators)

True Value	β_{12}	γ_{11}	γ_{12}
	.5	.513	.513
d_1d_2	-.7071	.1414	.1414
d_1d_3	5.5154*	.4243	4.9498*
d_1d_4	1.2728	-2.6870	-2.6870
d_1d_5	-.1414	.9899	1.2728
d_1d_6	5.2326*	.4243	5.5154
d_1d_7	3.8184*	.7071	-6.9297*
d_1d_8	3.8184*	-.7071	1.2728
d_1d_9	4.9498*	.4243	5.5154*
d_2d_3	5.5154*	.1414	.4243
d_2d_4	.9899	-.4243	-.4243
d_2d_5	.4243	.7071	-.4243
d_2d_6	5.2326*	-.1414	-.4243
d_2d_7	3.8184*	-.1414	-1.2728
d_2d_8	3.8184*	1.2728	-.9899
d_2d_9	4.9498*	.7071	.4243
d_3d_4	-5.2326*	-.4243	-6.0811*
d_3d_5	-5.2326*	.1414	.7071
d_3d_6	-.7071	-3.2527	2.6870*
d_3d_7	-1.2728	-.4243	-6.9297*
d_3d_8	-1.5556	.1414	.1414
d_3d_9	-3.2527*	5.5154*	.7071
d_4d_5	-.1414	.9899	1.5556
d_4d_6	5.2326*	.4243	6.3640*
d_4d_7	4.1012*	1.5556	-6.9297*
d_4d_8	3.8183*	.9899	1.2728
d_4d_9	4.9498*	.4243	6.3640*
d_5d_6	5.2326*	.4243	.4243
d_5d_7	3.8184*	-.9899	-1.5556
d_5d_8	4.1012*	-.9899	-.9899
d_5d_9	4.1012*	-.1414	-.7071
d_6d_7	-.9899	-.4243	-6.9297*
d_6d_8	-1.5556	.1414	.1414
d_6d_9	-3.2527*	6.9297*	-6.6468*
d_7d_8	-.4243	.7071	1.2728
d_7d_9	.7071	.4243	6.9297*
d_8d_9	1.5556	.1414	-.1414

* = significant at 5% level.

A positive Z-score indicates the first estimator is better than the second, the reverse is true for a negative number.

and is estimated by the proportion of cases in which $|d_i-\theta|$ exceeds $|d_j-\theta|$.

The power-efficiency of this "sign" test is about 2/3, a small price to pay for the freedom to leave unspecified the exact functional form of the exact distributions of the estimators. The test statistics are given in Table 6.4.3.

Based on all the information obtained, the following conclusion could be made:-

1. For the coefficients of the first structural equation, the conventional 2SLS (d_2) attains minimum MSE for the coefficient for y_{2t} individually but has almost the largest TMSE, next only to the estimator with ridge regression at the first stage alone (d_8). It is noted that modifications of 2SLS at the first stage alone, whether by PC (d_5) or RR (d_8) will have a TMSE not much different from that of no modification at all and may sometimes be worse - agreeing with the theoretical results in Section 6.2. (Note: TMSE is used here as an approximate measure only, even if they are meaningless for estimators whose finite sample second moment does not exist).

2. Strangely enough, if modification is done at the second stage alone (such as the Swamy and Metha ridge-type 2SLS) (d_3) gains in TMSE could be realised. Principal Components Approach appears to be able to gain dramatic reduction on TMSE when it is applied to the second stage, irrespective of what is used in the first stage $(d_1, d_4$ and $d_7)$. Applying PC at both stages (d_4) have the practical advantage of solving the undersized sample problem as well as risk reduction.

3. Applying Ridge Regression at the second stage $(d_3, d_6$ and $d_9)$ could also secure reduction in TMSE as against to no modification at the second stage. With Ridge Regression at both stages (d_9), the estimator

has the advantage of being consistent and asymptotically efficient (equivalent to 2SLS) and also possesses finite sample moments. Although, the Principal Component Approach managed to attain much more reduction in risk, it does not yield a consistent estimator in general and its result may lack any economic justification.

4. The TMSE of various estimators appear to be quite indifferent to the first stage modification but is very sensitive to the second stage modification. If one insists on consistency and small total mean square error, then applying ridge regression at both stages is a practical and satisfactory alternative to the conventional 2SLS, especially in a situation of undersized sample together with multicollinearity.

5. The results for the coefficients of the second reduced form equation need little comments as they correspond to what are expected in theory and in the Monte Carlo studies in Chapter 4 of this thesis. They are presented here to show that the ridge regression are a very reasonable choice if one's aim is to obtain estimators with small total mean square error.

6. For individual coefficients in the first structural equation, Summer's 'sign' test indicated that d_2 (the 2SLS) beats *almost* all other estimators significantly for the coefficients of the included endogenous variable. The margin is not significant in comparison with d_1 (the PCLS), d_4 (the PCPC) and d_5 (the LSPC). For the other coefficients, d_4 (the PCPC) and d_7 (the PCRR) appear to be estimators dominating the others.

7. Regarding the bias term, the conventional 2SLS remains to be relatively less biased than all other methods. Of course, in the reduced

form, LS is an unbiased procedure and the bias reported in Table 6.4.2 are due to sampling fluctuation only.

8. The effect of ridge method is felt most by the bias terms. The bias due to a second stage ridge appears to be more serious than a first stage. It is possibly because that the SSCBC criterion usually leads to a larger ridge parameter than necessary and hence causes excessive bias in the ridge estimates.

§6.5 Conclusion

This chapter aims at attacking the data problem frequently occur in the estimation of structural equation in large or medium size econometric models. The undersized sample problem and the problem of multicollinearity are defined and alternative solutions are suggested and their finite sample propeties are analysed. The sensitivity of the conventional 2SLS to multicollinearity has been brought up in various studies in the literature but relatively little has been suggested to tackle the problem. Even if some feasible modifications have been mentioned, the theoretical consequence of so doing in a small sample situation is unknown. We have been able to throw some light on the advantages and disadvantages of using various modifications at each of the two stages of the estimation procedures and point out the benefit of an educated use of ridge regression techniques in these circumstances. Obviously, our result is far from being conclusive and further investigation and actually implementing the techniques suggested herein would help economists to gain more knowledge about estimating economic relations.

Appendix to Chapter 6

Here we present a derivation of the expression for bias and mean square error of the GRIV estimator substantially along the lines of Kadane (1971).

Following Nagar (1959) it is assumed that the reduced form disturbances

$$\bar{V}_z = [V_1 : 0] \tag{A.1}$$

can be decomposed into two independent components u and W such that

$$\bar{V}_z = \underset{\sim}{u}\underset{\sim}{q}' + W \qquad \mathcal{E}u = 0 \qquad \mathcal{E}W = 0 \tag{A.2}$$

where

$$\underset{\sim}{q} = \frac{1}{T}\,\mathcal{E}(\bar{V}_z'u) \ . \tag{A.3}$$

Let

$$C_1 = \underset{\sim}{q}\underset{\sim}{q}' \qquad \mathcal{E}(uu') = I \tag{A.4}$$

$$C_2 = \frac{1}{T}\,\mathcal{E}(W'W) \tag{A.5}$$

$$\bar{S} = \bar{Z}'N\bar{V}_z + \bar{V}_z'N\bar{Z} \ . \tag{A.6}$$

(For convenience, we have used Z for Z_1 throughout).

For constant matrices A, R_1 and R_2, the following lemmas can be established

LEMMA 1. $\mathcal{E}(W'AW) = (\text{tr } A)C_2$

LEMMA 2. $\mathcal{E}(WAW') = (\text{tr } C_2A)I$

LEMMA 3. $\mathcal{E}(WAW) = A'C_2$

LEMMA 4. $\mathcal{E}(uu'Auu') = (\text{tr } A)I + 2A$

LEMMA 5. $\mathcal{E}(u'R_1uu'R_2u) = (\text{tr } R_1)(\text{tr } R_2) + 2\,\text{tr}\,(R_1R_2)$

LEMMA 6. $\mathcal{E}\bar{V}_zR_1uu'R_2\bar{V}_z = [(\text{tr } R_2)R_1' + 2R_2'R_1']C_1 + R_2'R_1'C_2$

LEMMA 7. $\mathcal{E}\bar{V}_z'R_1uu'R_2\bar{V}_z = [(\text{tr } R_1)(\text{tr } R_2) + 2(\text{tr } R_1R_2)]C_1 + (\text{tr } R_1R_2)C_2$

LEMMA 8. $\mathcal{E}\bar{V}_zR_1uu'R_2\bar{V}_z' = [(\text{tr } R_2C_1R_1)I_n + 2R_2C_1R_1] + (\text{tr } C_2R_1R_2)I_n$

LEMMA 9. $\mathcal{E}\bar{V}_z'R_1uu'R_2\bar{V}_z' = C_1[(\text{tr } R_1)R_2' + 2R_2'R_1'] + C_2R_2'R_1'$

LEMMA 10.
$$\begin{aligned}\mathcal{E}\bar{S}R_1uu'R_2\bar{S} &= \bar{Z}'N\{[(\text{tr } R_2\bar{Z}'N)R_1' + 2N\bar{Z}R_2'R_1]C_1 + N\bar{Z}R_2'R_1'C_2\} \\ &+ \{[(\text{tr } N\bar{Z}R_1)(\text{tr } R_2\bar{Z}'N) + 2(\text{tr } N\bar{Z}R_1R_2\bar{Z}'N)]C_1 + (\text{tr } N\bar{Z}R_1R_2\bar{Z}'N)C_2\} \\ &+ \bar{Z}'N\{[(\text{tr } R_2C_1R_1)I_n + 2R_2C_1R_1] + (\text{tr } C_2R_1R_2)I_n\}N\bar{Z} \\ &+ \{C_1[(\text{tr } N\bar{Z}R_1)R_2' + 2R_2'R_1'\bar{Z}'N] + C_2R_2'R_1'\bar{Z}'N\}N\bar{Z}\end{aligned}$$

LEMMA 11. $\mathcal{E}\bar{V}_z'R_1\bar{V}_zR_2uu' = C_1R_2[(\text{tr } R_1)I_n + 2R_1] + (\text{tr } R_1)C_2R_2$

LEMMA 12. $\mathcal{E}\bar{V}_zR_1\bar{V}_zR_2uu' = [(\text{tr } R_1'C_1R_2)I_n + 2R_1'C_1R_2] + R_1'C_2R_2$

LEMMA 13. $\mathcal{E}\bar{V}_zR_1\bar{V}_z'R_2uu' = (\text{tr } C_1R_1)[(\text{tr } R_2)I_n + 2R_2] + (\text{tr } C_2R_1)R_2$

LEMMA 14. $\mathcal{E}\bar{V}_z'R_1\bar{V}_z'R_2uu' = C_1R_1'[(\text{tr } R_2)I_n + 2R_2] + C_2R_1'R_2$

LEMMA 15.
$$\begin{aligned}\mathcal{E}\bar{S}R_1\bar{S}R_2uu' &= \bar{Z}'N\{[(\text{tr } N\bar{Z}R_1'C_1R_2)I_n + 2N\bar{Z}R_1'C_1R_2] + N\bar{Z}R_1'C_2R_2\} \\ &+ \bar{Z}'N\{(\text{tr } C_1R_1)[(\text{tr } N\bar{Z}R_2)I_n + 2N\bar{Z}R_2] + (\text{tr } C_2R_1)N\bar{Z}R_2\} \\ &+ \{C_1R_2[(\text{tr } N\bar{Z}R_1\bar{Z}'N)I_n + 2N\bar{Z}R_1\bar{Z}'N] + (\text{tr } N\bar{Z}R_1\bar{Z}'N)C_2R_2\} \\ &+ \{C_1R_1'\bar{Z}'N[(\text{tr } N\bar{Z}R_2)I_n + 2N\bar{Z}R_2] + C_2R_1'\bar{Z}'N^2\bar{Z}R_2\} .\end{aligned}$$

Continuing with the notations in Section 6.2, the estimator error of GRIV is

$$e = \tilde{\delta} - \delta = (\tilde{Z}'Z)^{-1}\tilde{Z}'\sigma u \tag{A.7}$$

$$= \sigma[(N\bar{Z} + \sigma N\bar{V}_z + \Omega)'(\bar{Z} + \sigma\bar{V}_z)]^{-1}(N\bar{Z} + \sigma N\bar{V}_z + \Omega)'u \tag{A.8}$$

Terms inside square brackets in (A.8) can be written as

$$\begin{aligned}&(N\bar{Z} + \Omega)'\bar{Z} + \sigma(N\bar{Z} + \Omega)'\bar{V}_z + \sigma\bar{V}_z'N\bar{Z} + \sigma^2\bar{V}_z'N\bar{V}_z \\ &= Q^{-1} + \sigma\bar{S} + \sigma^2\bar{V}_z'N\bar{V}_z \\ &= Q^{-1}[I + Q\Delta]\end{aligned} \tag{A.9}$$

where

$$Q^{-1} = (N\bar{Z} + \Omega)'\bar{Z} = R'\bar{Z}$$

$$R = N\bar{Z} + \Omega$$

$$\Delta = \sigma\bar{S} + \sigma^2\bar{V}_z'N\bar{V}_z$$

$\Omega'\bar{V}_z \equiv 0$ and N is symmetric.

Substituting the geometric expansion of $(I+Q\Delta)^{-1}$ in (A.8) we obtain

$$e = \sigma[I - Q\Delta + Q\Delta Q\Delta - \ldots]Q(R'u + \sigma\bar{V}_z'Nu) \tag{A.10}$$

i.e.

$$Q^{-1}e = \sigma[I - \Delta Q + \Delta Q\Delta Q - \ldots](R'u + \sigma\bar{V}_z'Nu) \tag{A.11}$$

where

$$\Delta Q = \sigma\bar{S}Q + \sigma^2\bar{V}_z'N\bar{V}_zQ \tag{A.12}$$

$$\Delta Q\Delta Q = \sigma^2\bar{S}Q\bar{S}Q + \sigma^3(\bar{S}Q\bar{V}_z'N\bar{V}_zQ + \bar{V}_z'N\bar{V}_zQ\bar{S}Q) + \sigma^4\bar{V}_z'N\bar{V}_zQ\bar{V}_z'N\bar{V}_zQ\ . \tag{A.13}$$

So that

$$Q^{-1}e = \sigma[R'u + \bar{V}_z'Nu] - \sigma^2\bar{S}Q[R'u+\sigma\bar{V}_z'Nu] + \sigma^3[\bar{S}Q\bar{S}Q-\bar{V}_z'N\bar{V}_zQ]R'u + O(\sigma^4) \tag{A.14}$$

$$= \sigma B_1u + \sigma^2B_2u + \sigma^3B_3u + O(\sigma^4) \tag{A.15}$$

where

$$B_1 = R'$$

$$B_2 = \bar{V}_z'N - \bar{S}QR'$$

$$B_3 = \bar{S}Q\bar{S}QR' - \bar{V}_z'N\bar{V}_zQR' - \bar{S}Q\bar{V}_z'N\ .$$

The bias of $\tilde{\delta}$ is simply the expected value of e and we consider the expectations of the right-side of (A.14) term by term up to $O(\sigma^2)$:

$$\mathcal{E}B_1u = 0$$

$$\mathcal{E}B_2u = \mathcal{E}\bar{V}_z'Nu - \mathcal{E}\bar{S}QR'u$$

$$= \underset{\sim}{q}(\text{tr } N) + \bar{Z}NRQ\underset{\sim}{q} + (\text{tr } N\bar{Z}QR')\underset{\sim}{q}\ .$$

Consequently

$$\mathcal{E}(e) = \sigma^2[(\text{tr } N)Q - Q\bar{Z}'NRQ - (\text{tr } N\bar{Z}QR')Q]\underset{\sim}{q} + O(\sigma^3) \tag{A.16}$$

Q.E.D. □

This establishes Theorem 6.1 in Section 6.2.

We now give the derivation of the epxression for the mean square error matrix. Write

$$Q^{-1}ee'Q^{-1} = \sigma^2 B_1uu'B_1' + \sigma^3[B_1uu'B_2' + B_2uu'B_1'] + \sigma^4[B_1uu'B_3' + B_2uu'B_2' + B_3uu'B_1'] + O(\sigma^5). \qquad (A.16)$$

Again, taking expectation of the right side of (A.16) term by term up to $O(\sigma^4)$, we have

$$\mathcal{E}B_1uu'B_1' = \mathcal{E}R'uu'R = R'R$$

$$\mathcal{E}B_2uu'B_1' = \mathcal{E}(\bar{V}_z'N - \bar{S}QR')uu'R = 0 = B_1uu'B_2'$$

$$\mathcal{E}B_2uu'B_2' = \mathcal{E}(\bar{V}_z'N - \bar{S}QR')uu'(N\bar{V}_z - RQ\bar{S})$$

(i) $\mathcal{E}\bar{V}_z'Nuu'N\bar{V}_z = [(\text{tr } N)^2 + 2(\text{tr } N^2)]C_1 + (\text{tr } N^2)C_2$ (by Lemma 7 $R_1=R_2=N$)

(ii) $\mathcal{E}\bar{V}_z'Nuu'RQ\bar{S} = \mathcal{E}\bar{V}_z'Nuu'RQ(\bar{V}_z'N\bar{Z} + \bar{Z}'N\bar{V}_z)$

$$= C_1[(\text{tr } N)QR' + 2QR'N] + C_2QR'N'N\bar{Z} + [(\text{tr } N)(\text{tr } RQ\bar{Z}'N) + 2(\text{tr } NRQ\bar{Z}'N)]C_1 + (\text{tr } NRQ\bar{Z}'N)C_2$$

(by Lemma 9 $R_1=N$, $R_2=RQ$ and by Lemma 7 $R_1=N$, $R_2=RQ\bar{Z}'N$)

(iii) $\mathcal{E}\bar{S}QR'uu'N\bar{V}_z$ = transpose of (ii)

(iv) $\mathcal{E}\bar{S}QR'uu'RQ\bar{S} = \bar{Z}'N\{[(\text{tr } RQ\bar{Z}'N)RQ + 2N\bar{Z}QR'RQ]C_1 + N\bar{Z}QR'RQC_2\}$

$$+ \{[(\text{tr } N\bar{Z}QR')(\text{tr } RQ\bar{Z}'N) + 2(\text{tr } N\bar{Z}QR'RQ\bar{Z}'N)]C_1 + (\text{tr } N\bar{Z}QR'RQ\bar{Z}'N)C_2\}$$

$$+ \bar{Z}'N\{[(\text{tr } RQC_1QR')I_n + 2RQC_1QR'] + (\text{tr } C_2QR'RQ)I_n\}N\bar{Z}$$

$$+ \{C_1[(\text{tr } N\bar{Z}QR')QR' + 2QR'RQ\bar{Z}'N] + C_2QR'RQ\bar{Z}'N\}N\bar{Z}$$

(by Lemma 10 $R_1=QR'$, $R_2=RQ$)

If $\Omega = 0$ and $N^2 = N + O(\sigma^2)$ so that $R = N\bar{Z}$, then

$$R'\bar{Z} = Q = R'R + O(\sigma^2)$$

we have, collecting the four expressions, in this special case

$$\begin{aligned}\&B_2uu'B_2' &= (\text{tr } QC_1)Q^{-1} + (\text{tr } QC_2)Q^{-1} \\ &+ \{(\text{tr } N)^2 + 2 \text{ tr } N^2 - 2(\text{tr } N+2)(n+1) + (n+2)^2 + 2\}C_1 \\ &+ (\text{tr } N^2 - n)C_2 + O(\sigma^2) \ .\end{aligned}$$

Furthermore,

$$\&B_1uu'B_3' = \&R'uu'[RQ\bar{S}Q\bar{S} - RQ\bar{V}_z'N\bar{V}_z - N\bar{V}_z'Q\bar{S}]$$

(i) $$\begin{aligned}\&R'uu'RQ\bar{S}Q\bar{S} &= R'\{(\text{tr } RQC_1QR')I_n + 2RQC_1QR' + RQC_2QR'\}R \\ &+ R'\{(\text{tr } C_1Q)[(\text{tr } RQR')I_n + 2RQR'] + (\text{tr } C_2Q)RQR'\}R \\ &+ R'\{[(\text{tr } RQR')I_n + 2RQR']RQC_1 + (\text{tr } RQR')RQC_2\} \\ &+ R'\{[(\text{tr } RQR')I_n + 2RQR']RQC_1 + RQR'NRQC_2\}\end{aligned}$$

(by Lemma 15 $R_1=Q$, $R_2=QR'$)

(ii) $$\&R'uu'RQ\bar{V}_z'N\bar{V}_z = R'\{[(\text{tr } N)I_n + 2N]RQC_1 + (\text{tr } N)RQC_2\}$$

(by Lemma 11 $R_1=N$, $R_2=QR'$)

(iii) $$\begin{aligned}\&R'uu'N\bar{V}_z'QS &= \&R'uu'N\bar{V}_zQ(\bar{Z}'N\bar{V}_z + \bar{V}'N\bar{Z}) \\ &= R'\{(\text{tr } C_1Q)[(\text{tr } N)I_n + 2N] + (\text{tr } C_2Q)\}N\bar{Z} \\ &+ R'\{[(\text{tr } N)I_n + 2N]N\bar{Z}QC_1 + N^2\bar{Z}QC_2\}\end{aligned}$$

(by Lemma 13 $R_1=Q$, $R_2=N$ and Lemma 14 $R_1=N\bar{Z}Q$, $R_2=N$) .

Again, specialising to the case $\Omega = 0$ and $N^2 = N + O(\sigma^2)$ and collecting the three expressions, we have

$$\&B_1uu'B_3' = (\text{tr } QC_1)Q^{-1}[n-\text{tr } N+1] + [2(n+1-\text{tr } N)]C_1 + (n+1-\text{tr } N)C_2 + O(\sigma^2)$$

$$\&B_3uu'B_1' = \text{transpose of } \&B_1uu'B_3' \ .$$

Consequently

$$\&Q^{-1}ee'Q^{-1} = \sigma^2Q^{-1} + \sigma^4(\bar{Z}'N_1\bar{Z}+A^*) + O(\sigma^5) \qquad (A.17)$$

where

$$N_1 = -X(X'X)^{-1}G\Lambda^2(\Lambda+\sigma^2\bar{K})^{-2}\bar{K}G'(X'X)^{-1}X' \quad \text{as} \quad K^* = \sigma^2\bar{K}$$

$$A^* = (\text{tr } QC_1)Q^{-1}(2n+3-2 \text{ tr } N) + (\text{tr } C_2Q)Q^{-1}$$
$$+ [(n+2-\text{tr } N)^2 + 2(\text{tr } N^2) - 2 \text{ tr } N+2]C_1$$
$$+ [\text{tr } N^2-2 \text{ tr } N+n+2]C_2$$

and establishes Theorem 2 of Section 6.2. Q.E.D. □

CHAPTER 7

SOME FURTHER RESULTS AND CONCLUSION

§7.1 Effect of Model Misspecification on the Ordinary Ridge Estimator

All previous analyses of ridge-type estimators were based on the assumption that the linear regression model is correctly specified. In empirical investigations, the researcher will rarely have any assurance that some variables have not been omitted or that some irrelevant variables have not been included. Furthermore, the problem of serial correlation or heteroscedasticity might have also been overlooked. Effects of such common misspecifications generally add to undermine the efficiency of OLS (see Chapter 2). Since the ridge estimator is designed to improve estimation, it would be instructive to examine the robustness of the ordinary ridge estimator (ORE) under various forms of misspecification.

(A) Disturbance Term Misspecification

To begin with, we shall consider the case of a model being correctly specified except for its error structure. That is, the true model is

$$y = X\beta + \varepsilon \tag{7.1}$$

$$\mathcal{E}\varepsilon = 0 \quad \text{and} \quad \mathcal{E}\varepsilon\varepsilon' = \sigma^2\Omega \ . \tag{7.2}$$

But we have wrongly assumed that Ω to be the identity matrix,[1] then the OLS estimator for β in (7.1) has a covariance matrix

$$\text{Var}(\hat{\beta}) = \sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1} \tag{7.3}$$

---

[1] This case is indeed very general as other non-identity matrix can be reduced to the identity matrix by linear transformation. [See Watson (1955)].

which exceeds the lower bound

$$\sigma^2(X'\Omega^{-1}X)^{-1} \tag{7.4}$$

by a positive semidefinite matrix. This lower bound is also the risk matrix measure of the best linear unbiased estimator (BLUE) for β. We note that the ordinary ridge estimator (ORE)

$$\tilde{\beta} = (X'X+kI)^{-1}X'y = Z_k\hat{\beta} \tag{7.5}$$

where

$$Z_k = (X'X+kI)^{-1}X'X$$

$$\hat{\beta} = (X'X)^{-1}X'y$$

has a risk matrix measure equal to

$$MSE(\tilde{\beta}) = \sigma^2(X'X+kI)^{-1}X'\Omega X(X'X+kI)^{-1} + (Z_k-I)\beta\beta'(Z_k-I)' \ . \tag{7.6}$$

Thus, the following result can be easily established.

Theorem 7.1.1.

There exists $k > 0$ such that the risk matrix measure of the BLUE for model (7.1) exceeds that of the ORE by a positive semidefinite matrix.

Proof.

$$\begin{aligned}
&\sigma^2(X'\Omega^{-1}X)^{-1} - MSE(\tilde{\beta}) \\
&= \sigma^2(X'X+kI)^{-1}\{(X'X+kI)(X'\Omega^{-1}X)^{-1}(X'X+kI) - X'\Omega X\}(X'X+kI)^{-1} \\
&\qquad - k^2(X'X+kI)^{-1}\beta\beta'(X'X+kI)^{-1}
\end{aligned} \tag{7.7}$$

which is positive semidefinite iff

$$k^2\beta'[(X'X+kI)(X'\Omega^{-1}X)^{-1}(X'X+kI) - X'\Omega X]^{-1}\beta < \sigma^2 \ . \tag{7.8}$$

For two positive definite matrices A and B, we use $A \geq B$ to mean $A - B$ is positive semidefinite, then the following inequalities hold

$$\mu_1 X'X \geq X'\Omega X \geq \mu_T X'X \tag{7.9}$$

and

$$\frac{1}{\mu_T} X'X \geq X'\Omega^{-1}X \geq \frac{1}{\mu_1} X'X \tag{7.10}$$

where $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_T$ are ordered e-roots of Ω.

Hence, the expression inside curly bracket of (7.7) is bounded below by

$$\begin{aligned} \min\{(X'X+kI)(X'\Omega^{-1}X)^{-1}(X'X+kI)-X'\Omega X\} = \\ = (X'X+kI)\left(\frac{1}{\mu_T} X'X\right)^{-1}(X'X+kI)-\mu_1 X'X \\ = (\mu_T-\mu_1)X'X+\mu_T(2kI+k^2(X'X)^{-1}) \ . \end{aligned} \tag{7.11}$$

If $\mu_1 = \mu_2 = \ldots = \mu_T = 1$, then the first term vanishes and the sufficient condition for ORE to dominate BLUE is the standard one given by Farebrother (1976) at (3.33).

In the non-standard case where $\Omega \neq I$, we necessarily have $\mu_T < \mu_1$ so that the above lower-bound matrix may be negative definite. To ensure positive semi-definiteness, we consider the further lower bound of (7.11)

$$(\mu_T-\mu_1)\lambda_1 I + \mu_T(2k+\frac{k^2}{\lambda_1}) I \tag{7.12}$$

because $\lambda_1 I \geq X'X \geq \lambda_p I$ and $\frac{1}{\lambda_p} I \geq (X'X)^{-1} \geq \frac{1}{\lambda_1} I$ *where* $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ are ordered e-roots of $X'X$.

It is then clear that sufficient conditions for ORE to dominate BLUE are

(i) $\mu_T(\lambda_1+k)^2 \geq \lambda_1^2\mu_1$

and

(ii) $|\beta|^2 \leq \frac{\sigma^2}{\lambda_1}[\mu_T(\lambda_1+k)^2 - \lambda_1^2\mu_1]$.

Q.E.D.

Remarks. This condition is certainly more stringent than the strong condition of Farebrother (1976). Condition (i) is necessary to ensure positive-definiteness of the matrix in question.

(7.12) becomes a very stringent condition.

Next, we consider the consequence of the two other common types of misspecification on the bias and variance of ORE. The problem was first considered by Trivedi (1979) who found that overspecification generally worsens both the bias and variance of the ORE whereas the effect of underspecification is not so clear-cut. Here, we found that underspecification does *not* necessarily worsen the mean square error property of ORE. Let us consider the models

$$y = X_1\beta_1 + \varepsilon \tag{7.13}$$

and

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon \tag{7.14}$$

where $\mathcal{E}\varepsilon = 0$ and $\mathcal{E}\varepsilon\varepsilon' = \sigma^2 I$. X_1 and X_2 are $(T\times p_1)$ and $(T\times p_2)$ matrices of exogenous variables. β_1 and β_2 are respectively $p_1\times 1$ and $p_2\times 1$ vectors of unknown parameters.

(B) Underspecification

Assume that, (7.14) is taken as the true model, so that the *proper* ORE is given by the solution of the normal equation

$$\begin{pmatrix} A_{11}+kI & A_{12} \\ A_{21} & A_{22}+kI \end{pmatrix}\begin{pmatrix} \tilde{\beta}_1^* \\ \tilde{\beta}_2^* \end{pmatrix} = \begin{pmatrix} X_1'y \\ X_2'y \end{pmatrix} \tag{7.15}$$

where $A_{11} = X_1'X_1 \quad A_{22} = X_2'X_2 \quad A_{12} = X_1'X_2 = A_{21}' \quad k \geq 0$.

Solving (7.15) for $\tilde{\beta}_1^*$ we have

$$\tilde{\beta}_1^* = [(A_{11}+kI)-A_{12}(A_{22}+kI)^{-1}A_{21}]^{-1}[X_1'y-A_{12}(A_{22}+kI)^{-1}X_2'y] \tag{7.16}$$

and

$$\mathcal{E}\tilde{\beta}_1^* = [(A_{11}+kI)-A_{12}(A_{22}+kI)^{-1}A_{21}]^{-1}[A_{11}\beta_1+A_{12}\beta_2-A_{12}(A_{22}+kI)^{-1}(A_{21}\beta_1+A_{22}\beta_2)]$$

$$= \beta_1 + k[(A_{11}+kI) - A_{12}(A_{22}+kI)^{-1}A_{21}]^{-1}[A_{12}(A_{22}+kI)^{-1}\beta_2 - \beta_1] \ . \qquad (7.17)$$

On the other hand, the ORE for the *underspecified* model (7.13) is given by

$$\tilde{\beta}_1^u = (A_{11}+kI)^{-1}X_1'y \qquad (7.18)$$

with mean

$$\begin{aligned} \mathcal{E}\tilde{\beta}_1^u &= (A_{11}+kI)^{-1}(A_{11}\beta_1 + A_{12}\beta_2) \\ &= \beta_1 + (A_{11}+kI)^{-1}[A_{12}\beta_2 - k\beta_1] \ . \end{aligned} \qquad (7.19)$$

From (7.17) and (7.19), it can be seen that the bias of the two ORE's would generally be different unless $A_{12} = 0$ (i.e. when the omitted variables are orthogonal to the included ones).

The variance of $\tilde{\beta}_1^*$ is given by the top left $p_1 \times p_1$ submatrix of

$$\sigma^2 \begin{pmatrix} A_{11}+kI & A_{12} \\ A_{21} & A_{22}+kI \end{pmatrix}^{-1} \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} A_{11}+kI & A_{12} \\ A_{21} & A_{22}+kI \end{pmatrix}^{-1} . \qquad (7.20)$$

Again, if $A_{12} = 0$, then it is easy to see that

$$\mathrm{var}(\tilde{\beta}_1^*) = \sigma^2(A_{11}+kI)^{-1}A_{11}(A_{11}+kI)^{-1} = \mathrm{var}(\tilde{\beta}_1^u) \ . \qquad (7.21)$$

One question that would naturally arise is whether or not under-specification and $A_{12} \neq 0$ necessarily worsens the mean square error property of the ORE. The answer is in the negative as will become clear in the ensuing discussion. Let us consider the limiting case when $k \to 0$, the bias and variance of the ordinary ridge estimator as $k \to 0$ are respectively

$$\lim_{k\to 0}(\mathcal{E}\tilde{\beta}_1^* - \beta_1) = 0$$

and

$$\lim_{k\to 0}(\mathcal{E}\tilde{\beta}_1^u - \beta_1) = A_{11}^{-1}A_{12}\beta_2$$

which are the well known results for the ordinary least squares.

On the other hand

$$\lim_{k\ 0} \text{var}(\tilde{\beta}_1^*) = \sigma^2(A_{11} - A_{12}A_{22}^{-1}A_{12})^{-1}$$

and

$$\lim_{k\to 0} \text{var}(\tilde{\beta}_1^u) = \sigma^2 A_{11}^{-1} .$$

Combining the results, we see that as $k \to 0$,

$$[\text{MSE}(\tilde{\beta}_1^*) - \text{MSE}(\tilde{\beta}_1^u)] = \sigma^2(A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1} - \sigma^2 A_{11}^{-1} - A_{11}^{-1}A_{12}\beta_2\beta_2'A_{21}A_{11}^{-1}$$

which is positive definite iff

$$\beta_2'A_{21}[A_{11}(A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1}A_{11} - A_{11}]^{-1}A_{12}\beta_2 \leq \sigma^2 . \qquad (7.22)$$

That is, if k is sufficiently small and β_2 is not too large, then it is possible that $\tilde{\beta}_1^u$ would dominate $\tilde{\beta}_1^*$. Condition (7.22) is in fact nothing more than the Toro-Vizcarrondo and Wallace condition for restricted least squares to dominate the unrestricted one. The other extreme can be easily handled. Since both $\tilde{\beta}_1^*$ and $\tilde{\beta}_1^u$ becomes the null vector as k approaches infinity. Both estimators have the same risk of $\beta_1\beta_1'$ in the limit.

(C) Overspecification

The converse of the above is a case where (7.13) is the true model so that the *proper* ORE is

$$\tilde{\beta}_1^{**} = (A_{11} + kI)^{-1}X_1'y \qquad (7.23)$$

with

$$\&\tilde{\beta}_1^{**} = (A_{11} + kI)^{-1}A_{11}\beta_1$$

$$= \beta_1 - k(A_{11} + kI)^{-1}\beta_1 . \qquad (7.24)$$

If we have used ORE for the *overspecified* model (7.14), we shall obtain

$$\tilde{\beta}_1^o = [(A_{11}+kI)-A_{12}(A_{22}+kI)^{-1}A_{21}]^{-1}[X_1'y-A_{12}(A_{22}+kI)^{-1}X_2'y] \tag{7.25}$$

with

$$\begin{aligned}\mathcal{E}\tilde{\beta}_1^o &= [(A_{11}+kI)-A_{12}(A_{22}+kI)^{-1}A_{21}]^{-1}[A_{11}\beta_1-A_{12}(A_{22}+kI)^{-1}A_{21}\beta_1] \\ &= \beta_1 - k[(A_{11}+kI) - A_{12}(A_{22}+kI)^{-1}A_{21}]^{-1}\beta_1 \ .\end{aligned} \tag{7.26}$$

As before, when the wrongly included variables are orthogonal to the core variables, then both ORE's have the same bias, namely

$$-k(A_{11}+kI)^{-1}\beta_1 \ . \tag{7.27}$$

Otherwise, the average bias of $\tilde{\beta}_1^o$ would be larger than that of $\tilde{\beta}_1^{**}$ because

$$k^2\beta_1[(A_{11}+kI)-A_{12}(A_{22}+kI)^{-1}A_{21}]^{-2}\beta_1 > k^2\beta_1(A_{11}+kI)^{-2}\beta_1 \tag{7.28}$$

which follows from the fact that the difference of the two positive definite matrices

$$\begin{aligned}&(A_{11}+kI) - [(A_{11}+kI) - A_{12}(A_{22}+kI)^{-1}A_{21}] \\ &\qquad = A_{12}(A_{22}+kI)^{-1}A_{21}\end{aligned} \tag{7.29}$$

is a positive semidefinite matrix.

Note that as k approaches zero both bias terms vanish, which conforms with the well known results for OLS in the case of overspecified models.

Turning to the variance matrices we see immediately from (7.23) that

$$\text{var } \tilde{\beta}_1^{**} = \sigma^2(A_{11}+kI)^{-1}A_{11}(A_{11}+kI)^{-1} \tag{7.30}$$

and from (7.25) that

$$\text{var } \beta_1^o = \sigma^2[(A_{11}+kI)-A_{12}(A_{22}+kI)^{-1}A_{21}]^{-1}(A_{11}-A_{12}(A_{22}+kI)^{-1}A_{21}$$
$$-kA_{12}(A_{22}+kI)^{-2}A_{21})[(A_{11}+kI)-A_{12}(A_{22}+kI)^{-1}A_{21}]^{-1} . \qquad (7.31)$$

Direct comparison of (7.30) and (7.31) appears to be difficult. Considering the limiting cases, we have

$$\lim_{k\to 0} \text{var } \tilde{\beta}_1^{**} = \sigma^2 A_{11}^{-1}$$

$$\lim_{k\to 0} \text{var } \tilde{\beta}_1^{o} = \sigma^2 (A_{11}-A_{12}A_{22}^{-1}A_{21})^{-1} .$$

As $A_{12}A_{22}^{-1}A_{21}$ is positive semidefinite, it follows that

$$(A_{11}-A_{12}A_{22}^{-1}A_{21})^{-1} - A_{11}^{-1} \qquad (7.32)$$

is positive semidefinite.

That means that for sufficiently small k, $\tilde{\beta}_1^o$ has a larger variance matrix than $\tilde{\beta}_1^{**}$. Together with the fact that $\tilde{\beta}_1^o$ has at least as large a bias as $\tilde{\beta}_1^{**}$, one can conclude that the risk matrix measure of $\tilde{\beta}_1^o$ would in general exceed that of $\tilde{\beta}_1^{**}$ by a positive semidefinite matrix (for a sufficiently small k). Note also that both $\tilde{\beta}_1^o$ and $\tilde{\beta}_1^{**}$ approaches the null vector as k approaches infinity, they are thus having the same mean square error in the limit.

The above discussions (subsections A, B and C) are restricted to what would have expected from the misspecifications on the mean square error properties of ridge estimator *in the long run*. One would naturally ask how does misspecification show up in the estimates. A general statement would be difficult to arrive at. However, some useful information would be obtained for certain data-dependent (adaptive) ridge estimator. Since there is a one to one correspondence between the ridge estimate and the ridge parameter, it is suffice to examine how misspecification

affects the determination of the ridge parameter. Among the various algorithms for estimating the ridge parameter, we have found that the one due to Lawless and Wang is particularly versatile and we shall base our discussion on that algorithm.

As is shown in Chapter 4, the Lawless-Wang ridge parameter for the p-variate regression model is given by

$$k_{LW} = \frac{p\hat{\sigma}^2}{\hat{\beta}'X'X\hat{\beta}} = \frac{p}{T-p}\,\frac{(1-R^2)}{R^2} \tag{7.33}$$

where R^2 = coefficient of determination.

Let R_i^2 $(i = A,B,C)$ denote the corresponding coefficient of determination under misspecifications of type A, B and C.

It is well known, from least squares theory, that

$$(1-R^2) \leq (1-R_i^2) \qquad i = A,B, \tag{7.34}$$

i.e. $$R^2 \geq R_i^2 .$$

For misspecification type A, we see that the factor $\frac{p}{T-p}$ remains unchanged but

$$\frac{1-R_A^2}{R_A^2} > \frac{1-R^2}{R^2} \tag{7.35}$$

which implies that

$$k_{LW}^A > k_{LW} . \tag{7.36}$$

That is, if autocorrelation or heteroscedasticity are overlooked, then it is certain that the Lawless-Wang type ridge parameter would be larger for the misspecified model than the one for the correct model. However, we know that $\text{var}(\hat{\beta}^A) > \text{var}(\hat{\beta})$ and that the expected length of $\hat{\beta}^A$ is longer than that of $\hat{\beta}$. It is likely that $\hat{\beta}^A$ is further

away from the true vector β than is $\hat{\beta}$. Hence a bigger shrinkage factor for $\hat{\beta}^A$ does not necessarily do more harm to the ridge estimates in this case.

On the other hand, if the model is underspecified then we can also establish similarly that

$$\frac{1-R_B^2}{R_B^2} > \frac{1-R^2}{R^2} \quad . \tag{7.37}$$

However, we also have

$$\frac{p}{T-p} > \frac{p_1}{T-p_1} \qquad T > p > p_1 > 0 \tag{7.38}$$

It is *not* so sure whether $k_{LW}^B > k_{LW}$ or $k_{LW} > k_{LW}^B$ in this case. Provided k_{LW}^B is small enough, the resulting ridge estimates will not be seriously affected by underspecification.

Finally, for the overspecified case, it is obvious that

$$\frac{1-R_C^2}{R_C^2} < \frac{1-R^2}{R^2} \tag{7.39}$$

but

$$\frac{p_1+p_2}{T-p_1-p_2} > \frac{p_1}{T-p_1} \quad . \tag{7.40}$$

Thus, even if the erroneously included variables are *orthogonal* to the core variables, the ridge estimates based on an overspecified model may be shrinked more to zero than the one based on a correct model. In general, the ridge estimator for an overspecified model may be more biased than that for the correct model.

We note in passing, that Teräsvirta (1979a) compared the Total Prediction Square Error (TPSE) of a slightly more general form of the ridge-type estimators with that of the OLS under the misspecifications

of type B and C. However, since our risk criterion is in terms of MSE-matrix which is stronger than the TPSE-criterion, our conditions for superiority are sufficient to ensure the conditions in his paper.

§7.2 Testing Stochastic Restrictions

The mixed estimator for the model

$$y = X\beta + u \qquad u \sim N_T(0,\sigma^2 I) \tag{7.41}$$

subject to the stochastic restrictions

$$r = R\beta + v \qquad v \sim N_m(\delta,\tau^2 I) \tag{7.42}$$

was given by Theil and Goldberger (1961) as

$$\bar{\beta} = (X'X + kR'R)^{-1}(X'y + kR'r) \tag{7.43}$$

where

$$k = \frac{\sigma^2}{\tau^2} \quad \text{is assumed to be given .} \tag{7.44}$$

Building on the work of Theil (1963), Yancey et al (1974) established that the compatibility statistic

$$\hat{\gamma} = \frac{(r-R\hat{\beta})'S_k(r-R\hat{\beta})}{m\hat{\sigma}^2} \tag{7.45}$$

has a *non-central* F-distribution on m and $T-p$ degrees of freedom with the non-centrality parameter

$$\lambda_k = \frac{\delta'S_k\delta}{\sigma^2} \tag{7.46}$$

where

$$S_k = [\tfrac{1}{k} I + R(X'X)^{-1}R']^{-1}$$

$$\hat{\sigma}^2 = y'(I-X(X'X)^{-1}X')y/(T-p) \ .$$

The statistic (7.45) has a *central* F-distribution under the null hypothesis $\mathcal{E}(r-R\beta) = \delta = 0$ if one is only interested in checking the compatibility between sample and prior information. However, if one's objective is to improve estimation in the sense of risk reduction, then a testing procedure analogous to those proposed by Wallace and his co-workers are desirable.

By reformulating the model with stochastic information as a model with exact restrictions, Fomby's (1978) test statistic arises naturally from the result of Toro-Vizcarrondo and Wallace (1968). Consider the combined model of (7.42) and (7.43)

$$\underset{(T+m)\times 1}{\begin{pmatrix} y \\ \sqrt{k}r \end{pmatrix}} = \begin{pmatrix} X & 0 \\ 0 & \sqrt{k}R \end{pmatrix} \beta + \begin{pmatrix} u \\ \sqrt{k}v \end{pmatrix} \quad \text{var} \begin{pmatrix} u \\ \sqrt{k}v \end{pmatrix} = \sigma^2 I_{T+m}$$

$$= \underset{(T+m)\times(p+m)}{\begin{pmatrix} X & 0 \\ 0 & \sqrt{k}I \end{pmatrix}} \underset{(p+m)\times 1}{\begin{pmatrix} \beta \\ \theta \end{pmatrix}} + \begin{pmatrix} u \\ \sqrt{k}\delta + w \end{pmatrix} \tag{7.47}$$

where

$$\sqrt{k}R\beta = \sqrt{k}\theta \qquad \mathcal{E}\begin{pmatrix} u \\ w \end{pmatrix} = 0 \qquad \text{var}\begin{pmatrix} u \\ w \end{pmatrix} = \sigma^2 I_{T+m}$$

which has a conventional form linear model as

$$y^* = Q\phi + \xi \qquad \xi' = (u', w') \tag{7.48}$$

with the exact linear restrictions

$$H\phi = h \tag{7.49}$$

where

$$y^* = \begin{pmatrix} y \\ \sqrt{k}r \end{pmatrix} \qquad Q = \begin{pmatrix} X & 0 \\ 0 & \sqrt{k}I \end{pmatrix} \qquad \phi = \begin{pmatrix} \beta \\ \theta + \sqrt{k}\delta \end{pmatrix}$$

$$\underset{m\times(p+m)}{H} = [R : -I] \qquad h = -\sqrt{k}\delta \ .$$

In this case, the restricted least square estimator of ϕ for the model (7.48) is

$$\tilde{\phi} = \hat{\phi} - (Q'Q)^{-1}H'[H(Q'Q)^{-1}H']^{-1}H\hat{\phi} \tag{7.50}$$

where

$$\hat{\phi} = (Q'Q)^{-1}Q'y^* \ .$$

The risk matrix measure of $\tilde{\phi}$ was exceeded by that of $\hat{\phi}$ by a positive semidefinite matrix iff the non-centrality parameter

$$\lambda_k = \frac{(H\phi-h)'[H(Q'Q)^{-1}H']^{-1}(H\phi-h)}{\sigma^2} \tag{7.51}$$

is less than unity.

This condition can be tested by the statistic

$$u_k = \frac{(H\hat{\phi}-h)'[H(Q'Q)^{-1}H']^{-1}(H\hat{\phi}-h)}{m\hat{\sigma}^{*2}} \tag{7.52}$$

where

$$\sigma^{*2} = \frac{y^{*\prime}(I-Q(Q'Q)^{-1}Q')y^*}{(T+m)-(p+m)}$$

u_k is known to have the non-central F-distribution on m and $T-p$ degrees of freedom with the non-centrality λ_k given at (7.51).

In the process of hypothesis testing, one can also compute u_k for successive values of $k \in [o,\infty)$. Since this is a continuously increasing function of k, one would expect that the null hypothesis of $\lambda_k = 1$ could be rejected for a high enough value of k at a pre-assigned level of significance. Table of non-central F-distribution with $\lambda_k = 1$ was given by Wallace and Toro-Vizcarrondo (1969) and can be used for the purpose of evaluating the appropriateness of introducing stochastic information. [Note: Our definition of non-centrality parameter differs from Wallace's by a factor of 2].

On the other hand, focusing his attention on the weak criterion of total prediction error, Teräsvirta (1979b) established that

$$\text{TPSE}(\bar{\beta}) = \sigma^2 p - \text{tr}(\sigma^2 S_k R(X'X)^{-1}R' - S_k \delta\delta' S_k R(X'X)^{-1}R') \tag{7.53}$$

which is a function of k that it first decreases until the global minimum is reached and then starts to increase before "asymptotically" approaching the risk of the restricted least square estimator, viz.

$$\lim_{k\to\infty} \text{TPSE}(\bar{\beta}) = \text{TPSE}(\tilde{\beta}) = \sigma^2(p-m) + \delta'(R(X'X)^{-1}R')^{-1}\delta \ .$$

Differentiating (7.53) with respect to k, we have

$$\frac{d}{dk}\text{TPSE}(\bar{\beta}) = -\text{tr}\ \sigma^2 \left(\frac{dS_k}{dk}\right) R(X'X)^{-1}R' + \left(\frac{dS_k}{dk}\right)\delta\delta' S_k R(X'X)^{-1}R'$$

$$+ S_k\delta\delta'\left(\frac{dS_k}{dk}\right) R(X'X)^{-1}R' \tag{7.54}$$

where

$$\frac{dS_k}{dk} = \frac{d}{dk}\left(\frac{1}{k} I + R(X'X)^{-1}R'\right)^{-1}$$

$$= \frac{1}{k^2}\left(\frac{1}{k} I + R(X'X)^{-1}R'\right)^{-2} = \frac{1}{k^2} S_k^2 \ . \tag{7.55}$$

Substituting (7.55) into (7.54) and rearranging, we have

$$\frac{d}{dk}\text{TPSE}(\bar{\beta}) = -\frac{1}{k^2}\text{tr}\, S_k R(X'X)^{-1}R' S_k(\sigma^2 S_k^{-1} - 2\delta\delta')S_k \tag{7.56}$$

which is non-positive when

$$\sigma^2 S_k^{-1} - 2\delta\delta' \tag{7.57}$$

is positive semidefinite. Hence, a *necessary and sufficient* condition that $\text{TPSE}(\bar{\beta})$ is still decreasing as a function of k is when

$$\frac{1}{\sigma^2}\delta' S_k \delta \leq \frac{1}{2} \ . \tag{7.58}$$

The inequality of (7.58) approaches

$$\frac{1}{\sigma^2}\,\delta'(R(X'X)^{-1}R')\delta \leq \frac{1}{2} \tag{7.59}$$

as k approaches infinity. Teräsvirta suggested that one should first test if (7.59) is valid so that exact restricted least square estimator is likely to be superior. If the tests rejects this hypothesis, then one can use (7.45) to test the inequality at (7.58) for increasing values of $k \in [0,\infty)$ to locate the minimum of the total prediction square error function. Note that, conventionally, the restricted least square is TPSE-superior to OLS when

$$\frac{1}{\sigma^2}\,\delta'(R(X'X)^{-1}R')\delta < m \tag{7.60}$$

where m = number of exact restrictions imposed (see Wallace (1972))

which indicates that Teräsvirta's condition at (7.59) is overstringent. One important reservation about the above tests is that k was assumed to be non-stochastic throughout. If k is in some sense stochastic, such as those picked up from a ridge trace or estimated by maximum likelihood estimates, then the exact distribution of the test statistic is unknown. Nevertheless, the Monte Carlo results of Yancey et al (1974) indicate that the non-central F-distribution is a fairly accurate approximation for the distribution of such a statistic. Furthermore, the risk of the resulting estimator will be impaired by such preliminary tests.

§7.3 Confidence Regions Based on Biased Estimates

We have now entered into a presumably the most controversial area of statistical inference. There seems to be little objection among Statisticians to gauging the performance of a estimator by some quadratic loss functions and thus tolerate the notion of a biased estimator.

However, some of them may reject the relevance of biased estimator on the ground that one cannot make probabilistic statements about them in the "conventional" way. By "conventional", we mean the school of thought of Neyman and Pearson, which defines probabilities in terms of relative frequencies in repeatable experiments. Two other schools of thought, namely, the fiducial approach and the Bayesian approach can also be used to estimate the interval in which the true parameter should lie at a given level of confidence. Of course, solutions would generally differ among the three approaches except in some simple cases. It is therefore worth elaborating here the fundamental conceptual differences among the three approaches in order to finally make our own recommendation in the construction of "confidence" intervals for unknown parameters.

Basically, the conventional approach rests on the assumption that a single sufficient statistic for the unknown parameter exists. From the sampling distribution of the sufficient statistic a random interval can be constructed which, in the long run, will cover the unknown parameter by a desired proportion of times in repeated trials. For instance, there are a number of trials, each of which takes a sample of n observations from a population whose members are distributed independently and identically as normal with common mean μ and common variance σ^2. Conditional on σ^2, the unique sufficient statistic is the sample mean $\bar{x}$ which has a sampling distribution of normal with mean μ and variance $\frac{\sigma^2}{n}$. Then, one can construct an interval for the unknown parameter μ as

$$\left(\bar{x} - \frac{1.96\sigma}{\sqrt{n}} , \quad \bar{x} + \frac{1.96\sigma}{\sqrt{n}} \right) \tag{7.61}$$

which varies as varying $\bar{x}$'s are observed. After a sufficiently large number of trials, one is *almost certain* that 95% of the intervals so constructed would cover the unknown mean μ. In actual fact, the

repeated trials are only conceptual experiments and the probabilistic statement about μ is based on only one trial.

On the other hand, the fiducial approach of Fisher is based on the concept of a likelihood function. It is simply another way to look at the probability density function of a random variable. Continuing with the example above, the unique sufficient statistic of μ, $\bar{x}$, has a probability distribution

$$dF(\bar{x}|\mu,\sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} e^{-\frac{n}{2}(\frac{\bar{x}-\mu}{\sigma})^2} d\bar{x} . \qquad (7.62)$$

Then conditional on $\bar{x}$ and σ^2, the *likelihood* of μ when interpreted as a measure of the intensity of belief about μ can be written as

$$dF(\mu|\bar{x},\sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} e^{-\frac{n}{2}(\frac{\bar{x}-\mu}{\sigma})^2} d\mu \qquad (7.63)$$

and is called the *fiducial distribution* of the parameter μ.

In other words μ is now regarded as a "random" variable with mean $\bar{x}$ and variance $\frac{\sigma^2}{n}$, and one can then select an interval in which one "believes" that μ will fall at a given level of probability. The obvious choice is one that is shortest for a given level of probability, which leads to an interval centering at $\bar{x}$. Thus, the fiducial approach and the conventional approach coincide in this simple case. However, this does not necessarily have to be so in general. One simple counter-example where the two approaches may differ is in estimating the difference of means between two normal distributions having unequal variances (the well-known Fisher-Behren's problem). Note that the fiducial argument is specific to a particular sample whereas the conventional argument refers to a "population" of samples, although both are drawing conclusions about the unknown μ from only one given sample.

Finally, the Bayesian's argument appears to be more flexible and perhaps more convincing. As we have noted in Chapter 1 that the basic tool for Bayesian inference is the posterior distribution, which can be derived from a prior distribution and a likelihood function. Since the Bayesians define probabilities as a numerical measure of degree of belief rather than the relative frequencies of occurrence of events in repeatable trials, there is a problem of objectivity and uniqueness. However, this can be remedied if the researcher and his clients *agree* on a particular prior distribution - or a series of plausible priors - to work with, then there should be no confusion about the final outcome of the analysis. Jeffrey's diffuse prior was an attempt to win public acceptance of a particular way to represent "ignorance". As it is well known that the combination of a diffuse prior on μ and a normal likelihood function yields a posterior distribution which has exactly the same form as the fiducial distribution at (7.63). It is not surprising that all three approaches to the interval estimation of μ in this simple example end up with the same confidence interval. One distinct advantage of Bayesian approach is that it is capable of incorporating more realistic prior information about the unknown parameters than the other two approaches.

Suppose in the above example the sample consists of n independent observations of x, each has the same unknown distribution with mean μ and variance σ^2 on the range $[0,1]$. Central limit theorem assures us that as $n \to \infty$, the sample mean $\bar{x}$ will be distributed as normal with mean μ and variance $\frac{\sigma^2}{n}$ and conventional interval estimation theory asserts that

$$\Pr\left\{ \bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} \right\} = 0.95 \ . \qquad (7.64)$$

It is quite likely that this would lead to an absurd statement such as

$$\Pr\{-1 \leq \mu \leq 2\} = 0.95 \tag{7.65}$$

whereas we have already known that $0 \leq \mu \leq 1$. A Bayesian approach which legitimately assumes a uniform prior distribution of μ on the range $(0,1)$ will avoid this embarrassment.

As all the biased estimators we have developed so far have a Bayesian justification, there is in principle no problem in constructing confidence regions if one accepts the principle of Bayesian inference. However, it is true that many researchers would feel uneasy with Bayesian arguments and several attempts have been made to find "better" conventional confidence regions based on the biased estimates. Stein (1962, 1974) first attempted to construct more powerful confidence sets for the m-mean problem but his results are rather incomplete. The corresponding biased confidence regions can be of various shapes and can have smaller volume than the conventional Minimum Variance Unbiased confidence region. Morris (1977) defined the variance of a biased estimate as the "estimated" posterior variance. The basic argument is that of an empirical Bayes. Consider the estimation of the m-mean $\underset{\sim}{\theta} = (\theta_1 \ldots \theta_m)'$ from the observed data

$$\underset{\sim}{x} \sim N_m(\underset{\sim}{\theta}, \sigma^2 I) \tag{7.66}$$

and assume that the true parameters $\underset{\sim}{\theta}$ follow a normal prior distribution

$$\underset{\sim}{\theta} \sim N_m(\mu \underset{\sim}{\ell}, \tau^2 I) \ . \tag{7.67}$$

where

$\underset{\sim}{\ell}$ is the m-vector of units.

If μ and τ^2 are known, the Bayes estimator of $\underset{\sim}{\theta}$ under squared error loss is the posterior mean

$$\mathcal{E}(\underset{\sim}{\theta} | \mu, \tau^2, \sigma^2, x) = \mu \underset{\sim}{\ell} + (1-a)(x - \mu \underset{\sim}{\ell})$$

where

$$a = \frac{\sigma^2}{\sigma^2 + \tau^2} .$$

and a posterior variance equals $(\frac{1}{\sigma^2} + \frac{1}{\tau^2})^{-1}$.

i.e.

$$\underset{\sim}{\theta} | \mu, \tau^2, \sigma^2, \underset{\sim}{x} \sim N_m(\mu \underset{\sim}{\ell} + (1-a)(\underset{\sim}{x} - \mu \underset{\sim}{\ell}),\ \sigma^2(1-a)I) . \tag{7.68}$$

It is noted that the posterior distribution of θ at (7.68) depends on a number of nuisance parameters, especially those in the prior distribution, namely, μ, τ^2. Since μ may assume any value on the real line, a diffuse prior for μ is a reasonable choice. Regarding (7.68) as the likelihood function for μ, then integrating out μ on $(-\infty, \infty)$ yields

$$\underset{\sim}{\theta} \Big|_{\tau^2, \sigma^2, \underset{\sim}{x}} \sim N(\bar{x} \underset{\sim}{\ell} + (1-a)(\underset{\sim}{x} - \bar{x} \underset{\sim}{\ell}),\ \sigma^2(1-a)I + \frac{a\sigma^2}{m} J) \tag{7.69}$$

where

$$J = \underset{\sim\sim}{\ell\ell}' .$$

This is similar to a result obtained through the hierarchical priors of Lindley and Smith (1972) except that here we still have to remove the nuisance parameter τ^2. However, the mean and variance at (7.69) is unknown if a is not known even if values of σ^2 and $\underset{\sim}{x}$ are given.

Since we have the "marginal" distribution for x as

$$x \Big|_{\mu, \tau^2, \sigma^2} \sim N_m(\mu \underset{\sim}{\ell}, (\sigma^2 + \tau^2)I) \tag{7.70}$$

it follows that

$$as = \frac{\Sigma(x_i - \bar{x})^2}{\sigma^2 + \tau^2} \sim \chi^2_{m-1} \tag{7.71}$$

where

$$s = \frac{1}{\sigma^2}\Sigma(x_i-\bar{x})^2 \ ,$$

it can easily be established that

$$\mathcal{E}\frac{m-3}{s} = (m-3)\,\mathcal{E}\frac{\sigma^2}{\Sigma(x_i-\bar{x})^2}$$

$$= (m-3)a\,\mathcal{E}\frac{1}{\chi^2_{m-1}}$$

$$= a \quad \text{(using Lemma 2.6.4).}$$

That is, $\frac{m-3}{s}$ provides an unbiased estimator for a, which can be substituted into the posterior distribution at (7.69). By so doing, the estimated mean of the posterior distribution at (7.69) becomes the famous James-Stein estimator. Such procedures are known as 'empirical' Bayes because the parameters of the prior distribution are obtained empirically rather than known a priori.

Morris (1977) observed that the James-Stein-type *empirical* Bayes posterior distribution assumes that τ^2 has a uniform distribution over $[-\sigma^2,\infty)$. An obvious improvement is to restrict τ^2 to be positive. Since

$$a = \frac{\sigma^2}{\sigma^2+\tau^2} \ , \text{ we have } \quad \tau^2 = \sigma^2(\frac{1}{a}-1) \tag{7.72}$$

and the Jacobian of transform from τ^2 to a is seen to be

$$\left|\frac{\sigma^2}{a^2}\right| \ . \tag{7.73}$$

From (7.71), we obtain the likelihood function for a, which is proportional to

$$dF(as) \propto e^{-\frac{as}{2}}(as)^{\frac{m-1}{2}-1}\,d(as) \tag{7.74}$$

while the prior distribution of a is

$$dG(a) \propto \frac{\sigma^2}{a^2}\, da \qquad 0 \leq a \leq 1 \ . \tag{7.75}$$

This is based on the assumption that τ^2 is uniform over $[0,\infty)$. Consequently, the posterior distribution of a given s is obtained from (7.74) and (7.75) as

$$dP(a|s) \propto e^{-\frac{as}{2}}\, a^{n-1}\, da \qquad 0 \leq a \leq 1 \tag{7.76}$$

where

$$n = \frac{m-3}{2} \ .$$

The formal Bayes estimate of a is thus given by

$$\tilde{a} = \mathcal{E}(a|s) = \frac{\int_0^1 a^n e^{-\frac{as}{2}}\, da}{\int_0^1 a^{n-1} e^{-\frac{as}{2}}\, da} \tag{7.77}$$

Integrating the numerator of (7.77) by parts once yield

$$-\frac{2}{s} e^{-\frac{s}{2}} + \frac{2n}{s} \int_0^1 a^{n-1}\, e^{-\frac{as}{2}}\, da \tag{7.78}$$

and hence (7.77) simplifies to

$$\tilde{a} = \frac{m-3}{s}\left[1 - \frac{1}{e_n(s)}\right] \tag{7.79}$$

$$= \hat{a}\left[1 - \frac{1}{e_n(s)}\right]$$

where

$$e_n(s) = ne^{\frac{s}{2}} \int_0^1 a^{n-1}\, e^{-\frac{as}{2}}\, da$$

$$n = \frac{m-3}{2}$$

$\hat{a} = \frac{m-3}{s}$ is the unbiased estimator of a.

Instead of the unbiased estimator $\hat{a}$, Morris used the formal Bayes estimator $\tilde{a}$ to *estimate* the unknowns in the posterior distribution of θ at (7.79). He conjectured that this new version of Empirical Bayes procedure would improve upon the James-Stein procedure. The empirical Bayes estimator for θ is then equal to the 'estimated' posterior mean

$$\mathcal{E}(\theta|x) = \bar{x}_{\underset{\sim}{\ell}} + (1-\tilde{a})(\underset{\sim}{x}-\bar{x}_{\underset{\sim}{\ell}}) \tag{7.80}$$

while the posterior variance can be obtained as

$$\begin{aligned} \mathrm{Var}(\theta|x) &= \mathcal{E}[(\mathrm{Var}(\theta|x,a)|x] + \mathrm{Var}[\mathcal{E}(\theta|x,a)|x] \\ &= \mathcal{E}[\sigma^2(1-a)I + \frac{a\sigma^2}{m} J|x] + \mathrm{Var}(a|x)(\underset{\sim}{x}-\bar{x}_{\underset{\sim}{\ell}})(\underset{\sim}{x}-\bar{x}_{\underset{\sim}{\ell}})' \\ &= \sigma^2(1-\tilde{a})I + \frac{\tilde{a}\sigma^2}{m} J + \sigma^2 v(\underset{\sim}{x}-\bar{x}_{\underset{\sim}{\ell}})(\underset{\sim}{x}-\bar{x}_{\underset{\sim}{\ell}})' \end{aligned} \tag{7.81}$$

where

$$\begin{aligned} v &= \mathrm{Var}(a|x) = -2\frac{d}{ds}\mathcal{E}(a|s) \\ &= \frac{2}{s}[\tilde{a} - n(1-\tilde{a})] \\ J &= \underset{\sim\sim}{\ell\ell}' \ . \end{aligned}$$

Finally, confidence regions about θ can be constructed using the estimated posterior distribution of θ given x, which is approximated by a multivariate normal with mean (7.80) and covariance matrix (7.81). The approximation is reasonable because had a been known a priori, the posterior distribution should be exactly normal.

Morris (1977) applied this technique both to real data and artificial data and found that $\mathcal{E}(\theta|x)$ at (7.80) attains a risk much smaller than of the maximum likelihood estimator $\bar{x}_{\underset{\sim}{\ell}}$ while confidence intervals constructed for individual components of θ using only diagonal elements of $\mathrm{var}(\theta|x)$ at (7.81) also provides much shorter confidence intervals than the conventional intervals and all of which are more conservative

in a sense that the intervals cover a higher than the nominal level of probability. The resulting intervals could not be accurate if the approximation to the posterior distribution by normal is not good and if the off-diagonal elements of $\text{var}(\theta|x)$ are non-negligibly large.

Vinod (1977) follows a slightly different approach to construct approximate confidence sets for the ridge estimators. The basic tool is Stein's identity.

It is well known that the regression model can be written in canonical form, using notations in Section 3.2 and Section 3.3

$$\begin{aligned} y &= X\beta + u \\ &= P\Lambda^{\frac{1}{2}}G'\beta + u \end{aligned} \tag{7.82}$$

i.e.

$$\Lambda^{-\frac{1}{2}}P'y = \gamma + \Lambda^{-\frac{1}{2}}P'u \tag{7.83}$$

where γ is the vector of principal components of β.

The OLS estimate of γ is

$$\hat{\gamma} \sim N_p(\gamma, \sigma^2\Lambda^{-1}) \ . \tag{7.84}$$

The generalised ridge estimator

$$\begin{aligned} \beta^* &= (X'X + GK^*G')^{-1}X'y \\ &= G\Delta\hat{\gamma} \end{aligned} \tag{7.85}$$

where

$$K^* = \text{diag}\{k_1 \ \ldots \ k_p\}$$

$$\Delta = \text{diag}\{\delta_1 \ldots \delta_p\} \qquad \delta_i = \frac{\lambda_i}{\lambda_i + k_i} \qquad i = 1 \ldots p \ .$$

When all k_i equal k, β^* becomes the ordinary ridge estimator which has been studied extensively in the literature.

Consider now the principal correlates between the Principal Co-ordinates of X and the response vector y.

$$\underset{\sim}{\hat{\theta}} = \frac{1}{\sigma} \Lambda^{\frac{1}{2}} \hat{\gamma} \sim N_p(\theta, I) \tag{7.86}$$

where

$$\theta = \frac{1}{\sigma} \Lambda^{\frac{1}{2}} \gamma .$$

That is, $\underset{\sim}{\hat{\theta}}$ are drawings from the standard p-mean model considered by Stein.

Let the biased estimator of θ_i be given by

$$\theta_i^* = \hat{\theta}_i + f(\hat{\theta}_i) \tag{7.87}$$

where $f(\cdot)$ is any continuous function of θ_i so that $\mathcal{E}\dot{f}(\hat{\theta}_i) < \infty$. Stein's identity states that

$$\mathcal{E}\dot{f}(\hat{\theta}_i) = \mathcal{E}[(\hat{\theta}_i - \theta_i) f(\hat{\theta}_i)] \tag{7.88}$$

where

$$\dot{f}(\hat{\theta}_i) = \frac{df(\hat{\theta}_i)}{d\hat{\theta}_i}$$

and the total mean square error of the biased estimator (7.87) equals (see Chapter 3)

$$\text{TMSE}[\hat{\theta}_i + f(\hat{\theta}_i)] = 1 + \mathcal{E}[f(\hat{\theta}_i)]^2 + 2\mathcal{E}[\dot{f}(\hat{\theta}_i)] . \tag{7.89}$$

Since both sides of (7.89) have the same expectation, the unbiased estimator of TMSE on the left side may be written in terms of known quantities on the right side when the expectation operates $\mathcal{E}$ are dropped. For the ordinary ridge estimator of θ_i with given values of δ_i, in $[0,1]$, we have

$$\delta_i \hat{\theta}_i = \hat{\theta}_i + (\delta_i - 1)\hat{\theta}_i . \tag{7.90}$$

The unbiased estimator of TMSE in (7.89) is seen to be

$$\hat{TMSE}(\delta_i\hat{\theta}_i) = 1 + (\delta_i-1)^2\hat{\theta}_i^2 + 2(\delta_i-1) \ . \qquad (7.91)$$

Thus, the shrinkage fraction δ_i will reduce the unbiased estimator of risk at (7.91) provided that

(i) $\delta_i < 1$

and

(ii) $2(\delta_i-1) > (\delta_i-1)^2\hat{\theta}_i^2$

i.e. $2 > (1-\delta_i)\hat{\gamma}_i^2\lambda_i/\sigma^2$.

Since treating k as nonstochastic, δ_i is nonstochastic and higher order derivatives of $f(\hat{\theta}_i)$ are all zero.

By repeated use of Stein's identity, Vinod obtained the following identity

$$\mathcal{E}(SD)^2 = \mathcal{E}\{[\delta_i\hat{\theta}_i-\theta_i]^2 - \hat{TMSE}(\delta_i\hat{\theta}_i)\}^2$$

$$= 4\delta_i^2 - 2 + 4(\delta_i-1)^2\mathcal{E}\hat{\theta}_i^2 \qquad (7.92)$$

An unbiased estimator of the (7.92) is, therefore, given by

$$(SD)^2 = 4\delta_i^2 - 2 + 4(\delta_i-1)^2\hat{\theta}_i^2 \ . \qquad (7.93)$$

Use the fact that, asymptotically,

$$\left[\frac{[\delta_i\hat{\theta}_i-\theta_i]^2 - \hat{TMSE}(\delta_i\hat{\theta}_i)}{SD}\right]^2 \sim \chi_1^2 \qquad (7.94)$$

so that approximately, from (7.91),

$$Pr\{(\delta_i\hat{\theta}_i-\theta_i)^2 < 1 + (\delta_i-1)^2\hat{\theta}_i^2 + 2(\delta_i-1) + c_\alpha(SD)\} = 1-\alpha \qquad (7.95)$$

where c_α is chosen so that

$$Pr\{\chi_1^2 < c_\alpha^2\} = 1-\alpha \ . \qquad (7.96)$$

Vinod proposes the *upper* and *lower* confidence limits for γ_i respectively as

$$\hat{\gamma}_i^{UP} = \delta_i \hat{\gamma}_i + \frac{\sigma}{\sqrt{\lambda_i}} [B + 2\delta_i - 1 + c_\alpha (SD)]^{\frac{1}{2}} \tag{7.97}$$

and

$$\hat{\gamma}_i^{LO} = \delta_i \hat{\gamma}_i + \frac{\sigma}{\sqrt{\lambda_i}} [B + 2\delta_i - 1 + c_\alpha (SD)]^{\frac{1}{2}} \tag{7.98}$$

where

$$\Pr[\hat{\gamma}_i^{LO} \leq \gamma_i \leq \hat{\gamma}_i^{UP}] = 1 - \alpha \qquad i = 1 \ldots p \, . \tag{7.99}$$

$$B = (\delta_i - 1)^2 \hat{\theta}_i^2 \, .$$

By definition, $\beta = G\gamma$ which implies $\beta_i = \Sigma g_{ij} \gamma_j$ where g_{ij} is the (i,j)'th element of G. A confidence interval for β_i will be estimated as

$$b_i^{UP} = \sum_{j=1}^{p} \max(g_{ij} \hat{\gamma}_j^{UP}, \, g_{ij} \hat{\gamma}_j^{LO})$$

and

$$b_i^{LO} = \sum_{j=1}^{p} \min(g_{ij} \hat{\gamma}_j^{UP}, \, g_{ij} \hat{\gamma}_j^{LO})$$

which is centered *near* the ridge estimator. Vinod claimed that this gives narrower interval reflecting a reduction in variance of the ridge estimator over the usual estimator. Note, the fact that δ_i is usually a stochastic quantity will seriously affect Vinod's result and the computation of the intervals appear to be unnecessarily complicated.

On the other hand, Obenchain (1977) advocated the use of the conventional confidence interval on the grounds that the *unbiased* confidence interval based upon the central F (or Student's t) distribution is shortest. Interval corresponding to a biased ridge estimate is identical to the least squares one if an unbiased estimate of the bias is used to correct the bias. Also, confidence regions which are shifted so as to be

centered at biased ridge estimators can have greater maximum diameter than does the least squares region of the same confidence. This criticism applies to both the Morris and Vinod intervals. Moreoever, results of Stein, Morris and Vinod provide "approximate" procedures based upon statistics with unknown nuisance parameters and their exact statistical properties are intractable. In view of these difficulties we suggest that either the conventional least square confidence interval or the empirical Bayesian posterior confidence interval should be used in connection with a risk-reducing biased estimate. More extensive studies of the techniques are needed before they can confidently be used.

§7.4 Critiques of Ridge Estimators

A major portion of this thesis is devoted to a study of the ridge estimator and its use in econometrics. However, this estimator has been subject to a number of well-founded criticisms which must be faced squarely by potential users of the technique. These criticisms are originated from both the classical school and the Bayesian school, and are directed principally at the ordinary ridge estimator of Hoerl and Kennard. To fix ideas, we list the major criticisms below and examine their justifications carefully one by one:-

(1) That the ridge estimates are not invariant to a nonsingular linear transformation of the linear model.

(2) That multicollinearity is just another way to describe the weak data problem so that a ridge estimator designed to solve collinearity problem does not necessarily lead to "better" parameter estimates because orthogonal data is not the same thing as strong data.

(3) The implicit prior information provided by the ridge estimator may not be compatible with the data and is no substitute for a proper Bayes estimator. Especially the shrinking of OLS to the origin is arbitrary.

(4) The statistical properties of a ridge estimator with a stochastically determined ridge parameter (k) are unknown. Even if k is nonstochastic, the estimator is biased and the bias depends on the unknown parameter.

(5) The effect of standardisation of data before application of ridge technique alters the loss function used to justify the estimator.

(6) The idea of coefficients stabilization or sign-correction is without scientific justification and could be misleading.

[See Conniffe and Stone (1973) or Campbell and Smith (1975)].

In what follows, we express our personal view about the criticisms. Regarding critique (1), it is true that the ridge estimates of β from the model

$$y = X\beta + u$$

will be different from $A^{-1}\theta^*$ where θ^* is the ridge estimator for $\theta = A\beta$ for the model

$$y = (XA^{-1})A\beta + u$$

whereas the OLS remains invariant. However, this does not invalidate the existence theorems about ridge estimators. That is, one can still be assured that OLS will be improved upon by a ridge estimator for the parameters of the *same* model. If one is interested in a linear transformation of the parameter, then one is in fact using a different model and a new ridge estimator can also be obtained to improve upon the OLS for the new model. It is rather unlikely that this invariance properties

of OLS in linear model will be of much practical importance in real data analysis, since if one is interested in marginal propensity to consume one would estimate it from the model directly and not from the marginal propensity to save.

Critique (2) is an extension of critique (1). Although any linear model can be transformed so that the explanatory variables are orthogonal, this has also reparameterised the model where the new parameters may not be of interest (especially in economics). As we have pointed out earlier in Chapter 3 that multicollinearity is not simply a weak data problem. It also causes difficulty in interpreting multidimensional evidence. Nevertheless, the application of ridge technique is susceptible to the same criticisms for the Generalised Inverse technique or truncation of Principal Components which correspond to small e-roots. All of which can be criticised as being ad hoc. The major difference is that the former introduces extra information while the latter throws away information. If one has reason to believe that the parameters are exchangeable and of small magnitudes, then a ridge estimator would be a sensible choice. Even if the exchangeable assumption does not hold, provided the signal to noise ratio is small, then a form of ridge estimator is superior to the usual estimator.

Critique (3) is fully justified and if the researcher is in possession of better prior information than the spherical normal prior centered at the origin, he should definitely be better off using a formal Bayes estimator than the ordinary ridge estimator. Theil-Goldberger's mixed estimator has already been noted as a valuable generalisation of the ridge technique. However, Vinod (1978) has shown that reduction in mean square error is still possible by shrinking the usual estimator to zero even if the true location is not the origin.

Critique (4) is valid in a limited sense. The rigorous analysis of a ridge estimator based on a stochastic k is always desirable, however, in its absence one can always rely on the experience from extensive Monte Carlo studies and real data analyses to identify situations where use of ridge estimator would be fruitful. If one is worried about the fact that the bias of a ridge estimate may be too big to realise any mean square error reduction over OLS, one may benefit from applying some of the testing procedures described in Section 7.2 to control the ridge estimator from being overbiased.

Critique (5) is valid but is not strictly indefensible. As the existence theorem states that there exists suitable ridge estimators which are better than OLS under a general quadratic loss function, one can be sure that this property will be preserved even after data standardisation. Nevertheless, the implied ridge estimates may be different from a direct ridge estimate unless one adopts the convention of applying ridge technique only to standardised data.

Finally, Critique (6) is unacceptable because sign-correction is *not* the major justification for ridge analysis. Nevertheless, in the process of stabilizing least square estimates, one is likely to have better point estimates - including one that has possibly the correct sign. Even if the sign may still be wrong, this does not mean that the total mean square error of the ridge estimator would not be smaller than that of the OLS. In practice, one usually has prior knowledge of the expected sign of a parameter. This information may help in obtaining better ridge estimates.

§7.5 Conclusion

We must confess in the conclusion of this chapter as well as of the thesis that there remain many unsolved problems relating to the use of biased estimators in econometrics. The more obvious unsolved problems are the mathematical analysis of the statistical distributions of non-linear biased estimators and their associated confidence intervals. Some of the philosophical problems relating to the incorporation of prior information are still unresolved and we do not see that they are likely to be solved in the near future. In the process of model building, economists are plagued by the poor quality of data that they typically possess. We have tried to identify some of the major defects of conventional approach to the estimation problem and to advocate throughout the thesis the careful use of all a priori information. The use of Bayesian methods may be helpful in obtaining better point estimates. However, one does not have to be a Bayesian to be able to take advantage of some of the techniques proposed in the thesis.

Although mechanistic application of the ordinary ridge estimator is not to be encouraged, we found that in many realistic situations, some version of the adaptive ridge estimator appear to perform very well in comparison with other biased and unbiased estimators. In suitable circumstances, their application is strongly recommended as the reduction in risk could be substantial. Two possible new areas in which the ridge estimator may be useful (but we have not entered into) are the estimation of translog production functions and the estimation of autoregressive time series models. This is because one would inevitably expect multi-collinearity among the explanatory variables in these models. It should also be noted that biased estimation techniques have been used long time ago in time series analysis to improve the estimate of the periodograms. One would also expect that a form of James-Stein estimator could be

gainfully employed in this situation.

No estimator is good for all models under all circumstances. Some educated choice must be made to select one that would minimise the risk pertinent to one's decision. We have reservation regarding the indiscriminate use of traditional least square estimators in econometric analysis. Nevertheless, they are important tools at the exploratory stage of model building and we argue strongly that in the final stage of an investigation, suitable biased estimates should also be computed and placed side by side with the usual estimates so that the decision maker will be able to select one that is compatible with his prior knowledge.

Throughout, we have surveyed and extended the existing literature on shrinkage estimation techniques and placed strong emphasis on the desirability of having an estimator that is admissible and minimax. The nature of econometric analysis is such that most of the good biased estimation procedure could not be applied to econometric models unless suitable modifications are made. Problems arising from using ridge estimators in Distributed Lag Models and Simultaneous Equation Models have been singled out for analysis in the thesis.

It must be pointed out that the ridge regression is based on the notion of correct specification of the model. It pays no attention to diagnostic checks which should be a part of any model building exercise. Although it is still possible to improve upon the usual estimator even if the ridge estimator were applied to a misspecified model, the amount of improvement will be impaired. That is why we recommend carrying out ridge analysis only after the conventional diagnostic procedures have been completed and careful thoughts have been given with respect to prior information to be incorporated.

We have not, however, dealt with the problems of structural changes and/or the need for robust estimation technique when the data distribution

departs from normal. Of course, if information about these non-standard situations is available, then further modification of the model is required before techniques proposed in this thesis can be gainfully employed.

BIBLIOGRAPHY

AITCHISON, J. and S.D. SILVEY (1958) Maximum Likelihood Estimation of Parameters Subject to Restraints, *Ann. Math. Statist.*, 29, 813-828.

ALAM, K. (1973) A Family of Admissible Minimax Estimators of the Mean of a Multivariate Normal Distribution, *Ann. Statist.*, 1, 517-525.

ALLEN, D.M. (1971) Mean Square Error of Prediction as a Criterion for Selecting Variables, *Technometrics*, 13, 469-475.

ALMON, S. (1965) The Distributed Lag Between Capital Approximations and Expenditures, *Econometrica*, 33, 178-196.

ALT, F.L. (1964) Distributed Lags, *Econometrica*, 10, 113-128.

AMEMIYA, T. (1966) On the Use of Principal Components of Independent Variables in Two Stage Least Squares Estimation, *Int'l Econ. Rev.*, 7, 283-303.

AMEMIYA, T. and W.A. FULLER (1967) A Comparative Study of Alternative Estimators in a Distributed Lag Model, *Econometrica*, 35, 509-529.

ANDERSON, T.W. (1971) *The Statistical Analysis of Time Series*, New York: Wiley.

ANDREWS, R.L., J.C. ARNOLD and R.G. KRUTCHKOFF (1972) Shrinkage of the Posterior Mean in the Normal Case, *Biometrika*, 59, 693-695.

BALDWIN, K.F. and A.E. HOERL (1978) Bounds on Minimum Mean Square Error in Ridge Regression, *Commun. in Statist.*, A7, 1209-1218.

BARANCHIK, A.J. (1970) A Family of Minimax Estimators of the Mean of a Multivariate Normal Distribution, *Ann. Math. Statist.*, 41, 642-645.

BARNARD, G.A. (1963) The Logic of Least Squares, *J. R. Statist. Soc. Series B*, 25, 124-127.

BHATTACHARYA, P.K. (1967) Estimates of a Probability Density Function and Its Derivatives, *Sankhya*, 29, 373-382.

BOCK, M.E. (1975) Minimax Estimators of the Mean of a Multivariate Normal Distribution, *Ann. Statist.*, 3, 209-218.

BOCK, M.E., T.A. YANCEY and G.G. JUDGE (1973) The Statistical Consequences of Preliminary Test Estimators in Regression, *J. Amer. Statist. Assoc.*, 68, 109-113.

BREUSCH, T.S. (1978) Some Aspects of Statistical Inference for Econometrics. Unpublished Ph.D. Dissertation, Australian National University.

BROOK, R.J. (1976) On the Use of a Regret Function to Set Significance Points in Prior Tests of Estimation, *J. Amer. Statist. Assoc.*, 71, 126-131.

BROWN, L.D. (1971) Admissible Estimator, Recurrent Diffusions and Insoluble Boundary Value Problems, *Ann. Math. Statist.*, 42, 855-903.

BROWN, P.J. and C. PAYNE (1975) Election Night Forecasting, *J. R. Statist. Soc. Series A*, 138, 463-498.

CAMPBELL, F. and G. SMITH (1975) A Critical Analysis of Ridge Regression, *Cowles Commission Discussion Paper.*

CARGILL, T.F. and R.A. MEYER (1974) Some Time and Frequency Domain Distributed Lag Estimators: A Comparative Monte Carlo Study, *Econometrica*, 42, 1031-1044.

CHIPMAN, J.S. (1964) On Least Squares with Insufficient Observations, *J. Amer. Statist. Assoc.*, 59, 1078-1111.

COHEN, A. (1965) Estimates of Linear Combinations of the Parameters in the Mean Vector of a Multivariate Distribution, *Ann. Math. Statist.*, 36, 78-87.

CONNIFFE, D. and J. STONE (1973) A Critical View of Ridge Regression, *The Statistician*, 22, 181-7.

DE FINETTI, B. (1974) *Theory of Probability*. New York: Wiley.

DEMPSTER, A.P., M. SCHATZOFF and N. WERMUTH (1977) A Simulation Study of Alternatives to Ordinary Least Squares, *J. Amer. Statist. Assoc.*, 72, 77-106.

DHRYMES, P.J. (1970) *Econometrics: Statistical Foundations and Applications*. New York: Harper and Row.

DHRYMES, P.J. (1971) *Distributed Lags*. San Francisco: Holden-Day.

DICKEY, J. (1974) Bayesian Alternatives to the F-test and Least Squares Estimate in the Normal Linear Model (reprinted in Fienberg and Zellner (1977) ed.) *Studies in Bayesian Econometrics and Statistics*, Vol. II, Ch. 12.1, 205-244.

DORAN, H.E. (1976) A Spectral Principal Components Estimator of the Distributed Lag Model, *Int'l Econ. Rev.*, 17, 8-25.

DREZE, J.H. (1977) Bayesian Regression Analysis Using Poly-t Densities, *J. of Econometrica*, 6, 329-354.

EFRON, B. and C. MORRIS (1973) Stein's Estimation Rule and Its Competitors - An Empirical Bayes Approach, *J. Amer. Statist. Assoc.*, 68, 117-130.

EFRON, B. and C. MORRIS (1976) Families of Minimax Estimators of the Mean of a Multivariate Normal Distribution, *Ann. Statist.*, 4, 11-12.

ENGLE, R.F. (1974) Specification of the Disturbance for Efficient Estimation, *Econometrica*, 42, 135-147.

ENGLE, R.F. and R. GARDNER (1976) Some Finite Sample Properties of Spectral Estimators of a Linear Regression, *Econometrica*, 44, 149-165.

FAITH, R.E. (1978) Minimax Bayes Estimators of a Multivariate Normal Mean, *J. of Multivariate Anal.*, 8, 372-379.

FAREBROTHER, R.W. (1976) Further Results on the Mean Square Error of Ridge Regression, *J. R. Statist. Soc. Series B*, 38, 248-250.

FARRAR, D.E. and R.R. GLAUBER (1967) Multicollinearity in Regression Analysis - The Problem Revisited, *Rev. of Econ. and Statist.*, 49, 92-107.

FISHER, F.M. (1965a) The Choice of Instrumental Variables in the Estimation of Economy-wide Econometric Models, *Int'l Econ. Rev.*, 6, 245-274.

FISHER, F.M. (1965b) Dynamic Structure and Estimation in Economy-wide Econometric Models, in J. Duesenberry, G. Fromm, L.R. Klein and E. Kuh (eds). *The Brookings Quarterly Model of the U.S.* Chicago: Rank-McNally, 589-636.

FISHER, I. (1937) Note on a Short-cut Method for Calculating Distributed Lags, *Bulletin de l'Institut International de Statistique*, 29, 323-328.

FISHMAN, G.S. (1969) *Spectral Methods in Econometrics*. Cambridge: Harvard University Press.

FOMBY, T.B. (1979) MSE Evaluation of Shiller's Smoothness Priors, *Int'l Econ. Rev.*, 20, 203-215.

FRISCH, R. (1934) *Statistical Confluence Analysis by Means of Complete Regression Systems*. Oslo: Universitetets Økonomiske Institutt.

GEARY, R.C. and C.E. LESER (1968) Significance Tests in Multiple Regression, *The Amer. Statistician*, 22, 20-21.

GERSOVITZ, M. and J.G. MacKINNON (1978) Seasonality in Regression: An Application of Smoothness Priors, *J. Amer. Statist. Assoc.*, 73, 264-273.

GOLDSTEIN, M. and A.F.M. SMITH (1974) Ridge-type Estimators for Regression Analysis, *J. R. Statist. Soc. Series B*, 36, 284-291.

GOODNIGHT, J. and T.D. WALLACE (1972) Operational Techniques and Tables for Making Weak MSE Tests for Restrictions in Regressions, *Econometrica*, 40, 699-710.

GUNST, R.F. and R.L. MASON (1977) Biased Estimation in Regression: An Evaluation Using Mean Squared Error, *J. Amer. Statist. Assoc.*, 72, 616-627.

HAAVELMO, T. (1950) Remarks on Frisch's Confluence Analysis and Its Use in Econometrics [in T.C. Koopmans ed., *Statistical Inference in Dynamic Economics*, New York: Wiley.]

HANNAN, E.J. (1965) The Estimation of Relationships Involving Distributed Lags, *Econometrica*, 33, 206-224.

HEMMERLE, W.J. (1975) An Explicit Solution for Generalised Ridge Regression, *Technometrics*, 17, 309-314.

HENDRY, D.F. and A.R. PAGAN (1979) Distributed Lags: A Survey of Some Recent Development, mimeo, Australian National University.

HILL, B.M. (1974) On Coherence, Inadmissibility and Inference About Many Parametrics in the Theory of Least Squares (reprinted in Fienberg and Zellner, ed.) *Studies in Bayesian Econometrics and Statistics*, Vol. II, Ch. 12.2, 245-274.

HOCKING, R.R., F.M. SPEED and M.T. LYNN (1976) A Class of Biased Estimators in Linear Regression, *Technometrics*, 18, 425-38.

HOERL, A.E. and R.W. KENNARD (1970) Ridge Regression: Applications to Non-Orthogonal Problems, *Technometrics*, 12, 69-82.

HOERL, A.E. and R.W. KENNARD (1970) Ridge Regression: Biased Estimation for Non-Orthogonal Problems, *Technometrics*, 12, 55-67.

HOERL, A.E. and R.W. KENNARD (1975) A Note on a Power Generalization of Ridge Regression, *Technometrics*, 17, 269.

HOERL, A.E. and R.W. KENNARD (1976) Ridge Regression: Iterative Estimation of the Biasing Parameter, *Commun. in Statist.*, A5, 76-88.

HOERL, A.E., R.W. KENNARD and K.F. BALDWIN (1975) Ridge Regression: Some Simulations, *Commun. in Statist.*, A4, 105-123.

HUNT, B.R. (1971) Biased Estimation for Non-Parametric Identification of Linear Systems, *Math. Biosci.*, 10, 215-237.

HUNTSBERGER, D.V. (1955) A Generalisation of a Preliminary Testing Procedure for Pooling Data, *Amer. Math. Statist.*, 26, 734-743.

JAMES, W. and C. STEIN (1961) Estimation with Quadratic Loss, *Proc. 4th Berkeley*, 361-379.

JEFFREYS, H. (1961) *Theory of Probability*. Oxford: Clarendon Press.

JOHNSTON, J. (1972) *Econometric Methods*, 2nd ed., New York: McGraw-Hill.

JORGENSON, D.W. (1963) Capital Theory and Investment Behaviour, *Amer. Econ. Rev.*, 53, 247-259.

JUDGE, G.G. and M.E. BOCK (1978) *The Statistical Implications of Pre-Test and Stein Rule Estimators in Econometrics*. Amsterdam: North-Holland.

KADANE, J.B. (1971) Comparison of k-class Estimators When the Disturbances are Small, *Econometrica*, 39, 723-737.

KLEIN, L.R. and M. NAKAMURA (1962) Singularity in the Equation System of Econometrics: Some Aspect of the Problem of Multicollinearity, *Int'l Econ. Rev.*, 3, 274-299.

KLOEK, T. and L.B.M. MENNES (1960) Simultaneous Equations Estimation Based on Principal Components of Predetermined Variable, *Econometrica*, 28, 45-61.

KNOTT, M. (1975) On the Minimum Efficiency of Least Squares, *Biometrika*, 62, 29-132.

KOYCK, L.M. (1954) *Distributed Lags and Investment Analysis.* Amsterdam: North-Holland.

LAWLESS, J.F. (1978) Ridge and Related Estimation Procedures: Theory and Practice, *Commun. in Statist.*, A7, 139-164.

LAWLESS, J.F. and P. WANG (1976) A Simulation Study of Ridge and Other Regression Estimators, *Coomun. in Statist.*, A5, 307-323.

LEAMER, E.E. (1969) Inference with Non-Experimental Data: A Bayesian View. Unpublished Ph.D. Dissertation, University of Michigan.

LEAMER, E.E. (1973) Multicollinearity: A Bayesian Interpretation, *Rev. of Econ. and Statist.*, 55, 371-80.

LEAMER, E.E. (1975) A Result on the Sign of Restricted Least Squares Estimates, *J. of Econometrics*, 3, 387-390.

LEAMER, E.E. (1977) Valley Regression: Biased Estimation for Orthogonal Problems, mimeo, University of California, Los Angeles.

LEAMER, E.E. (1978) *Specification Searches: Ad Hoc Inference With Non-Experimental Data.* New York: Wiley.

LEAMER, E.E. and G. CHAMBERLAIN (1976) A Bayesian Interpretation of Pretesting, *J. R. Statist. Soc. Series B*, 38, 85-94.

LEE, B.M.S. (1977) Justification for the Ridge Estimator for the Regression Coefficients in the General Linear Model, mimeo, Australian National University, Canberra.

LEE, B.M.S. (1978a) An Application of the Recursive Algorithm to the General Ridge Trace, mimeo, Australian National University, Canberra. (Submitted to *Technometrics*).

LEE, B.M.S. (1978b) Practical Ridge Analysis, mimeo, Australian National University, Canberra.

LEE, B.M.S. (1978c) Ridge-like Estimators for the Finite Distributed Lag Model, mimeo, Australian National University, Canberra.

LEE, B.M.S. (1979) Estimation of Linear Econometric Models Under Quadratic Loss, mimeo, Australian National University, Canberra.

LEE, B.M.S. and P.K. TRIVEDI (1979) Instrumental Variable Estimation of Structural Equations in Undersized Sample (submitted to *Econometrica*).

LESER, C.E. (1968) A Survey of Econometrics, *J. R. Statist. Soc. Series A*, 131, 530-566.

LINDLEY, D.V. and A.F.M. SMITH (1972) Bayes Estimates for the Linear Model, *J. R. Statist. Soc. Series B*, 34, 1-41.

LIU, T.C. (1960) Underidentification, Structural Estimation and Forecasting, *Econometrica*, 28, 855-865.

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

MALINVAUD, E. (1970) *Statistical Methods of Econometrics.* Amsterdam: North-Holland.

MALLOWS, C.L. (1973) Some Comments on C_p, *Technometrics,* 15, 661-675.

MARIANO, R.S. (1972) The Existence of Moments of the Ordinary Least Squares and Two Stage Least Squares Estimators, *Econometrica,* 40, 643-652.

MARQUARDT, D.W. (1970) Generalised Inverses, Ridge Regression, Biased Linear Estimation and Non-Linear Estimation, *Technometrics,* 12, 591-612.

MARTZ, H.F. and R.G. KRUTCHKOFF (1969) Empirical Bayes Estimators in a Multiple Linear Regression, *Biometrika,* 56, 367-74.

MAYER, L.S. and T.A. WILLKE (1973) On Biased Estimation in Linear Models, *Technometrics,* 15, 497-508.

McDONALD, G.C. and D.I. GALARNEAU (1975) A Monte Carlo Evaluation of Some Ridge Type Estimators, *J. Amer. Statist. Assoc.,* 70, 407-416.

MEHTA, J.S. and P.A.V.B. SWAMY (1978) The Existence of Moments of Some Simple Bayes Estimators of Coefficients in a Simultaneous Equation Model, *J. of Econometrics,* 7, 1-13.

MITCHELL, B. and F.M. FISHER (1970) The Choice of Instrumental Variables in the Estimation of Economy-wide Models: Some Further Thoughts, *Int'l Econ. Rev.,* 11, 226-234.

MIZON, G.E. (1977) Inferential Procedures in Nonlinear Models: An Application in a U.K. Industrial Cross Section Study of Factor Substitution and Returns to Scale, *Econometrica,* 45, 1221-1242.

MORRIS, C. (1977) Interval Estimation for Empirical Bayes Generalizations of Stein's Estimator. Unpublished Research Paper, Santa Monica: Rand Corporation.

MORRISON, J.L. (1970) Small Sample Properties of Selected Distributed Lag Estimators, *Int'l Econ. Rev.,* 11, 13-23.

MOSBAEK, E.J. and H.O. WOLD (1970) *Interdependent Systems.* Amsterdam: North-Holland.

MUTH, J.F. (1961) Rational Expectations and the Theory of Price Movements, *Econometrica,* 29, 315-335.

NEELEMAN, D. (1973) *Multicollinearity in Linear Economic Models.* The Netherlands: Tibury University Press.

NERLOVE, M. (1972) Lags in Economic Behaviour, *Econometrica,* 40, 221-252.

NEWHOUSE, J.P. (1971) Is Collinearity a Problem? *Rand Paper No.p-4588,* Santa Monica: Rand Corporation.

NEWHOUSE, J.P. and S.D. OMAN (1971) An Evaluation of Ridge Estimators, *Report No.R-716-PR.* Santa Monica: Rand Corporation.

NICHOLLS, D.F., A.R. PAGAN and R.D. TERRELL (1975) The Estimation and Use of Models with Moving Average Disturbance Terms: A Survey, *Int'l Econ. Rev.*, 16, 113-134.

OBENCHAIN, R.L. (1975) Ridge Analysis Following a Preliminary Test of a Shrunken Hypothesis, *Technometrics*, 17, 431-441.

OBENCHAIN, R.L. (1977) Classical F-tests and Confidence Regions for Ridge Regression, *Technometrics*, 19, 429-439.

PHILLIPS, D. (1962) A Technique for the Numerical Solution of Certain Integral Equations of the First Kind, *J. Assoc. Comput.*, 9, 97-101.

PLACKETT, R.L. (1950) Some Theorems in Least Squares, *Biometrika*, 37, 149-157.

RAIFFA, H. and R. SCHLAIFER (1961) *Applied Statistical Decision Theory.* Boston: Harvard Business School.

RAO, C.R. (1945) Generalization of Markoff's Theorem and Tests of Linear Hypotheses, *Sankhya*, 7 Part I, 9-16.

RAO, C.R. (1973) *Linear Statistical Inference and Its Applications*, 2nd ed., New York: Wiley.

RAO, C.R. (1975) Simultaneous Estimation of Parameters in Different Linear Models and Application to Biometric Problems, *Biometrics*, 31, 545-553.

RAO, C.R. (1976) Estimation of Parameter in a Linear Model, *Ann. Statist.*, 4, 1023-1037.

RICHARD, J.F. (1977) Bayesian Analysis of the Regression Model When the Disturbance are Generated by an Autoregressive Process (in A. Aykac and C. Brumat (eds) *New Developments in the Applications of Bayesian Methods*) p.185-209.

ROBBINS, H. (1955) An Empirical Bayes Approach to Statistics. *Proc. 3rd Berkeley Symposium in Math. Statistics*, Vol. I, 157-163. Berkeley: U.C.P.

ROTHENBERG, T.J. (1973) *Efficient Estimation with a Prior Information.* Monograph 23 Cowles Foundation.

SACKS, J. (1963) Generalized Bayes Solutions in Estimation Problems, *Ann. Math. Statist.*, 34, 751-768.

SARGAN, J.D. (1973) Model Building and Data Mining, L.S.E. Papers.

SARGAN, J.D. (1975) Asymptotic Theory and Large Models, *Int'l Econ. Rev.*, 16, 75-91.

SASTRY, M.V.R. (1970) Some Limits in the Theory of Multicollinearity, *The Amer. Statistician*, 24, 39-40.

SAWA, T. (1972) Finite Sample Properties of the k-class Estimators, *Econometrica*, 40, 653-680.

SAWA, T. and T. HIROMATSU (1973) Minimax Regret Significance Points for a Preliminary Test in Regression Analysis, *Econometrica*, 41, 1093-1102.

SCLOVE, S.L., C. MORRIS and R. RADHAKRISHNAN (1972) Non-Optimality of Preliminary Test Estimators for the Mean of a Multivariate Normal Distribution, *Ann. Math. Statist.*, 43, 1481-1490.

SHILLER, R.J. (1973) A Distributed Lag Estimator Derived From Smoothness Priors, *Econometrica*, 41, 775-788.

SILVEY, S.D. (1970) *Statistical Inference*. Baltimore: Penguin.

SINGH, R.S. (1977) Applications of Estimates of a Density and Its Derivatives to Certain Statistical Problems, *J. R. Statist. Soc., Series B*, 39, 357-363.

SIMS, C.A. (1974) Distributed Lags (in M.D. Intriligator and D.A. Kendrick (ed.): Frontiers of Quantitative Economics. Amsterdam: North Holland) Vol. 2, Ch.5.

STEIN, C. (1956) Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution, *Proc. 3rd Berkeley Symposium in Math. Statistics*, Vol. I, 197-206, Berkeley: U.C.P.

STEIN, C. (1962) Confidence Set for the Mean of a Multivariate Normal Distribution, *J. R. Statist. Soc. B*, 24, 265-296.

STEIN, C. (1974) Estimation of the Parameters of a Multivariate Normal Distribution I. Technical Report No.63, Department of Statistics, Stanford University.

STRAWDERMAN, W.E. (1971) Proper Bayes Minimax Estimators of the Multivariate Normal Mean, *Ann. Math. Statist.*, 42, 385-388.

STRAWDERMAN, W.E. (1972) On the Existence of Proper Bayes Minimax Estimators of the Mean of a Multivariate Normal Distribution, *Proc. 6th Berkeley Symposium of Math. Statistics*, Vol. I, 51-55, Berkeley: U.C.P.

STRAWDERMAN, W.E. (1978) Minimax Adaptive Generalised Ridge Regression Estimator, *J. Amer. Statist. Assoc.*, 73, 623-627.

STRAWDERMAN, W.E. and A. COHEN (1971) Admissibility of Estimators of the Mean Vector of a Multivariate Normal Distribution with Quadratic Loss, *Ann. Math. Statist.*, 42, 270-296.

SUMMERS, R. (1965) A Capital Intensive Approach to the Small Sample Properties of Various Simultaneous Equation Estimators, *Econometrica*, 33, 1-41.

SWAMY, P.A.V.B. and J. HOLMES (1971) The Use of Undersized Sample in the Estimation of Simultaneous Equation Systems, *Econometrica*, 39, 455-459.

SWAMY, P.A.V.B. and J.S. MEHTA (1976) Minimum Average Risk Estimators for Coefficients in Linear Models, *Commun. in Statist.*, A5, 803-818.

SWAMY, P.A.V.B., J.S. MEHTA and P.N. RAPPOPORT (1978) Two Methods of Evaluating Hoerl and Kennard's Ridge Regression, *Commun. in Statist.*, A7, 1133-1155.

SWINDEL, B.F. (1976) Good Ridge Estimators Based on Prior Information, *Commun. in Statist.*, A5, 1065-1075.

TERÄSVIRTA, T. (1979a) The Polynomial Distributed Lag Model, Discussion Paper No.7919. Belgium: C.O.R.E.

TERÄSVIRTA, T. (1979b) Some Results on Improving the Least Squares Estimation of Linear Models by Mixed Estimation. Discussion Paper No.7914. Belgium: C.O.R.E.

THEIL, H. (1961) *Economic Forecasts and Policy.* Amsterdam: North Holland.

THEIL, H. (1963) On the Use of Incomplete Prior Information in Regression Analysis, *J. Amer. Statist. Assoc.*, 58, 401-414.

THEIL, H. (1971) *Principles of Econometrics.* New York: Wiley.

THEIL, H. (1973) A Simple Modification of Two Stage Least Squares Procedures for Undersized Samples. In A.S. Goldberger and O.D. Duncan (eds) *Structural Equation Models in the Social Sciences,* New York: Seminar Press.

THEIL, H. and A.S. GOLDBERGER (1961) One Pure and Mixed Statistical Estimation in Economics, *Int'l Econ. Rev.*, 2, 65-78.

THEOBALD, C.M. (1974) Generalization of Mean Square Error Applied to Ridge Regression, *J. R. Statist. Soc. Series B,* 36, 103-6.

TIAO, G.C. and A. ZELLNER (1964) Bayes' Theorem and the Use of Prior Knowledge in Regression Analysis, *Biometrika,* 51, 219-230.

TINBERGEN, J. (1937) *An Econometric Approach to Business Cycle Problems.* Paris.

TORO-VIZCARRONDO, C.E. and T.D. WALLACE (1968) A Test of the MSE Criterion for Restrictions in Linear Regression, *J. Amer. Statist. Assoc.*, 63, 558-572.

TOYODA, T. and T.D. WALLACE (1976) Optimal Critical Values for Pre-Testing in Regression, *Econometrica,* 44, 365-375.

TRIVEDI, P.K. (1978) Estimation of a Distributed Lag Model Under Quadratic Loss, *Econometrica,* 46, 1181-1192.

TRIVEDI, P.K. (1979) Effects of Model Misspecification on Properties of the Simple Ridge Estimator, mimeo, Australian National University.

TRIVEDI, P.K. and B.M.S. LEE (1979) Seasonal Variability in Distributed Lag Models, mimeo, Australian National University.

TRIVEDI, P.K. and A.R. PAGAN (1979) Polynomial Distributed Lags: A Unified Treatment, *Econ. Studies Quarterly.*

VALENTINE, T.J. (1969) A Note on Multicollinearity, *Aust. Econ. Papers,* 8, 99-105.

VINOD, H.D. (1974) Ridge Estimation of a Trans-log Production Function, *Proc. Business & Econ. Statist. Section, Amer. Statist. Assoc.*, 596-601.

VINOD, H.D. (1976) Application of New Ridge Regression Methods to a Study of Bell System Scale Economics, *J. Amer. Statist. Assoc.*, 71, 835-841.

VINOD, H.D. (1977) Estimating the Largest Acceptable k and a Confidence Interval for Ridge Regression Coefficients. Paper presented at the Econometric Society European Meeting, Vienna.

VINOD, H.D. (1978a) Equivariance of Ridge Estimators Through Standardization, A Note, *Commun. in Statist.*, A7, 1157-1161.

VINOD, H.D. (1978b) A Ridge Estimator Whose MSE Dominate OLS, *Int'l Econ. Rev.*, 19, 727-737.

VINOD, H.D. (1978c) A Survey of Ridge Regression and Related Techniques for Improvement Over OLS, *Rev. of Econ. and Statist.*, 60, 121-131.

VISCO, I. (1978) On Obtaining the Right Sign of a Coefficient Estimate by Omitting a Variable from the Regression, *J. of Econometrics*, 6, 115-118.

WAHBA, G. (1976) A Survey of Some Smoothing Problems and the Method of Generalised Cross-Validation for Solving Them [In P.R. Krishnaiah, ed.: *Proc. of Symposium on the Application of Statistics*].

WALLACE, T.D. (1972) Weaker Criteria and Tests for Linear Restrictions in Regression, *Econometrica*, 40, 689-698.

WALLACE, T.D. and C.E. TORO-VIZCARRONDO (1969) Tables for the Mean Square Error Test for Exact Linear Restrictions in Regression, *J. Amer. Statist. Assoc.*, 64, 1649-1663.

WALLIS, K.F. (1967) Lagged Dependent Variables and Serially Correlated Errors: A Reappraisal of Three-Pass Least Squares, *Rev. Econ. Statist.*, 49, 555-567.

WALLIS, K.F. (1969) Some Recent Developments in Applied Econometrics: Dynamic Models and Simultaneous Equation Systems, *J. Econ. Lit.*, 7, 771-796.

WALLIS, K.F. (1977) Econometric Implications of Rational Expectation Hypothesis. Paper presented at the Econometric Society European Meeting, Vienna.

WATSON, G.S. (1955) Serial Correlation in Regression Analysis I, *Biometrika*, 42, 327-341.

WEBSTER, J.T., R.F. GUNST and R.L. MASON (1974) Latent Root Regression Analysis, *Technometrics*, 16, 513-22.

WICHERN, D.W. and G.A. CHURCHILL (1978) A Comparison of Ridge Estimators, *Technometrics*, 20, 301-310.

YANCEY, T.A., G.G. JUDGE and M.E. BOCK (1974) A Mean Square Error Test When Stochastic Restrictions are Used in Regression, *Commun. in Statist.*, A3, 755-768.

ZELLNER, A. (1971) *An Introduction to Bayesian Inference in Econometrics*, New York: Wiley.

ZELLNER, A. and W. VANDAELE (1975) Bayes-Stein Estimators for K-means, Regression and Simultaneous Equation Models. In: S. Fienberg and A. Zellner (eds) (1977). *Studies in Bayesian Econometrics and Statistics*, 629-654. Amsterdam: North Holland.

Supplementary References

BROWN, W.G. and B.R. BEATTIE (1975) Improving Estimates of Economic Parameters by Use of Ridge Regression with Production Function Application, *Amer. J. Agric. Econ.*, 57, 21-32.

CHRIST, C.F. (1966) *Econometric Models and Methods*, New York: Wiley.

DORAN, H.E. (1974) A Simulation Study of the Small Sample Properties of the Hannan Estimator of a Distributed Lag Model When the Signal-to-Noise Ratio is Constant, *Int'l Econ. Rev.*, 15, 497-513.

GUILKEY, D.K. and J.L. MURPHY (1975) Directed Ridge Regression Techniques in Cases of Multicollinearity, *J. Amer. Statist. Assoc.*, 70, 769-775.

HAITOVSKY, Y. and WAX, Y. (1976) Generalized Ridge Regression, Least Squares with Stochastic Prior Information and Bayesian Estimators. *Discussion Paper No.7626.* Belgium: C.O.R.E.

RICHMOND, J. (1974) Identifiability in Linear Models, *Econometrica*, 42, 731-736.

ROTHENBERG, T.J. (1971) Identification in Parametric Models, *Econometrica*, 39, 577-592.