

1988

Stein-Like Estimation and Inference.

Lee Chester Adkins

Louisiana State University and Agricultural & Mechanical College

Follow this and additional works at: https://digitalcommons.lsu.edu/gradschool_disstheses

Recommended Citation

Adkins, Lee Chester, "Stein-Like Estimation and Inference." (1988). *LSU Historical Dissertations and Theses*. 4552.

https://digitalcommons.lsu.edu/gradschool_disstheses/4552

This Dissertation is brought to you for free and open access by the Graduate School at LSU Digital Commons. It has been accepted for inclusion in LSU Historical Dissertations and Theses by an authorized administrator of LSU Digital Commons. For more information, please contact gradetd@lsu.edu.

INFORMATION TO USERS

The most advanced technology has been used to photograph and reproduce this manuscript from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book. These are also available as one exposure on a standard 35mm slide or as a 17" x 23" black and white photographic print for an additional charge.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.



University Microfilms International
A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
313/761-4700 800/521-0600

Order Number 8904521

Stein-like estimation and inference

Adkins, Lee Chester, Ph.D.

The Louisiana State University and Agricultural and Mechanical Col., 1988

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

Stein-Like Estimation and Inference

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The Department of Economics

by

Lee C. Adkins

A.A., Pensacola Junior College, 1978
B.S., Florida State University, 1980
M.S., Louisiana State University, 1985
August 1988

Contents

Chapter

1 Overview

1.1	Introduction	1
1.2	Out-of-Sample Forecasting With Stein- Rule Estimators	6
1.3	Improved Confidence Intervals and Ellipsoids	9
1.4	Risk Characteristics of a Stein-Like Estimator of the Probit Regression Model	13
1.5	Plan	14

2 Biased Estimation 16

2.1	Classical Linear Regression Model	18
2.2	Statistical Decision Theory	20
2.3	Alternative Estimation Rules	27
2.4	The Hypothesis Restricted Regression Model	41
2.5	Pretest Estimators	48
2.6	Stein-Rule Estimators	57
2.7	Stein-Rule Problems and Alternatives	63
2.8	Confidence Sets and Hypothesis Testing	66

3 Statistical Models and Methods 70

3.1	Time Series Models	71
3.2	Nonlinear Models	107
3.3	Generalized Linear Models	124
3.4	Computer Intensive Research Techniques	139

4	Improved Forecasts of Nominal GNP Growth	146
	Using the St. Louis Equation	
4.1	Introduction	147
4.2	The St. Louis Equation	150
4.3	The Statistical Model and Estimators . . .	153
4.4	Data Analysis	157
4.5	Least Squares Estimation	160
4.6	Members of the General Family of Minimax Estimators	165
4.7	ARIMA Forecasts	185
4.8	Results	188
4.9	Summary and Conclusion	194
A.4.1	Collinearity	209
A.4.2	Assessment of Similarity of In-Sample and Out-of-Sample Regressor Matrices . .	214
A.4.3	Regression Diagnostics	219
5	Improved Confidence Intervals and Ellipsoids . . .	229
	for the Linear Regression Model	
5.1	Interval Estimation and Hypothesis Testing Using Biased Estimators	230
5.2	Statistical Models and Data Generation Process	234
5.3	Estimators	235
5.4	The Bootstrap	247
5.5	Confidence Ellipsoids	266
5.6	Results	283
5.7	Conclusion	292

6	Risk Characteristics of a Stein-Like Estimator . .	298
	for the Probit Regression Model	
6.1	Introduction	299
6.2	Classical Normal Linear Regression Model and Estimators	302
6.3	The Probit Regression Model	306
6.4	Shrinkage Estimator for the Probit Regression Model	311
6.5	Data Generation	314
6.6	Results	315
6.7	Conclusion	318
7	Concluding Remarks	322
	References	329

Tables

4.1	OLS Estimates for the St. Louis Equation	161
4.2	Model Selection Estimates	174
4.3	Model Selection Hypothesis Restrictions for Principal Components	182
4.4	Hypothesis Restricted Estimators	184
4.5	Summary of Estimators and Their Symbols	185
4.6	ARIMA Diagnostics	187
4.7	RMSE Comparison Between RLS and OLS Estimators, 1962:2-1979:3	199
4.8	RMSE Comparison Between Stein-Rule and OLS Estimators, 1962:2-1979:3	200
4.9	RMSE Comparison Between Stein-Rule and RLS Estimators, 1962:2-1979:3	201
4.10	RMSE Comparison Between Stein-Rule and ARIMA, 1962:2-1982:3	202
4.11	RMSE Comparison Between RLS and OLS Estimators, 1962:2-1982:3	203
4.12	RMSE Comparison Between Stein-Rule and OLS Estimators, 1962:2-1982:3	204
4.13	RMSE Comparison Between Stein-Rule and RLS Estimators, 1962:2-1982:3	205
4.14	RMSE Comparison Between Stein-Rule and ARIMA, 1962:2-1982:3	206
A.4.1.a	Characteristic Roots and Condition Numbers, 1962:2-1979:3	212
A.4.1.b	Characteristic Roots and Condition Numbers, 1962:2-1982:3	213
A.4.2.a	Characteristic Roots for Centered In-Sample and Out-of-Sample Data (unscaled)	217
A.4.2.b	Characteristic Roots for Centered In-Sample and Out-of-Sample Data (scaled)	218
A.4.3.a	Tests for Nonnormality, Case II	223
A.4.3.b	Tests for Nonnormality, Case III	223

A.4.3.c	Tests for Nonnormality	223
A.4.3.d	Goldfeld-Quandt Tests	226
A.4.3.e	Breush-Pagan Tests	228
5.1	Summary Statistics, Monte Carlo	240
5.2	Averages of the Summary Statistics, Monte Carlo	242
5.3	Kolmogorov's D Statistics, Monte Carlo	247
5.4	Summary Statistics, Bootstrap	254
5.5	Averages of the Summary Statistics, Bootstrap	256
5.6	Percentile Confidence Intervals, 90%	261
5.7	Percentile Confidence Intervals, 95%	262
5.8	PRMSE of Estimated Covariance Matrices	273
5.9	Normal Approximation Intervals, 90%	286
5.10	Normal Approximation Intervals, 95%	287
5.11	Percentile Confidence Ellipsoids, 90%	291
5.12	Percentile Confidence Ellipsoids, 95%	291
6.1	Risk Characteristics, T=50	316
6.2	Risk Characteristics, T=100	318

Figures

4.1	Forecasted vs. Actual Values of GNP Growth, 1979:4-1983:3	207
4.2	Forecasted vs. Actual Values of GNP Growth, 1982:4-1986:3	208
5.1	Histogram of a Typical Least Squares Estimate From the Monte Carlo	243
5.2	Histogram of a Typical James-Stein Estimate From the Monte Carlo	244
5.3	Least Squares and James-Stein Histograms From the Monte Carlo	245
5.4	Histograms of Typical James-Stein Estimates From the Monte Carlo	246
5.5	Least Squares and James-Stein Histograms From the Bootstrap	263
5.6	Histogram of a Typical Least Squares Estimate From the Bootstrap	264
5.7	Histogram of a Typical James-Stein Estimate From the Bootstrap	265
5.8	Histograms of $Q_1^*(b)$ and $F(K, T-K)$ Random Variables	281
6.1	Risk of MLE, RMLE, Pretest, Stein, and Positive-Part Stein-Rule Estimators	320
6.2	Risk of MLE, RMLE, Pretest, Stein, and Positive-Part Stein-Rule Estimators	321

Abstract The dissertation addresses three issues in the use of Stein-like estimators of the classical normal linear regression model. The St. Louis equation is used to generate out-of-sample forecasts using least squares. These forecasts are compared to those produced by restricted least squares, pretest, and members from a general family of minimax shrinkage estimators using root-mean-square error criterion. Bootstrap confidence intervals and ellipsoids are constructed which are centered at least squares and James-Stein estimators and their coverage probability and size is explored in an Monte Carlo experiment. A Stein-like estimator of the probit regression model is suggested and its quadratic risk properties are explored in a Monte Carlo experiment.

CHAPTER 1 Overview

1.1 Introduction

Economic researchers have been notably reluctant to adopt new or improved statistical techniques. For example, in 1949 Cochrane and Orcutt proposed a "tentative" iterative procedure for eliminating autocorrelation in the linear regression model. Their technique -- often referred to as CORC -- requires the sacrifice of the first sample observation, a feature duly noted by the authors (1949, p. 59). Five years after the appearance of CORC, Prais and Winsten (1954) discovered a transformation which alleviated this troublesome feature of the Cochrane-Orcutt procedure. As Oxley and Roberts (1982) point out, many authors apparently refer to the Cochrane and Orcutt article as justification for the use of CORC without ever having read the paper or considered the clear misgivings it contains. The pitfalls of using CORC in small samples were recognized and an alternative was available; nevertheless, economists hesitated to adopt better techniques even after computational barriers fell.

It is no surprise, then, that economists have been reluctant to adopt more exotic and forbidding estimation procedures. For instance, biased estimators represent a large set of techniques yet to be assimilated into the mainstream of econometric practice. A biased estimator is one that yields, on average, parameter estimates which systematically differ from the true, but unknown population

parameters. The class of biased estimators is rapidly growing and includes such procedures as ridge regression [Hoerl and Kennard (1970a), (1970b)], adaptive ordinary ridge regression [Hoerl, Kennard and Baldwin (1975); Lawless and Wang (1976); McDonald and Galarneau (1975)]; RIDGM [Dempster, Schatzoff, and Wermuth (1977)], adaptive generalized ridge regression [Hemmerle and Brantle (1978); Strawderman (1978)], and pretest estimation [Toyoda and Wallace (1975); Ohtani and Toyoda (1980); Judge and Bock [(1978), Ch. 7]. In many situations biased estimators may be 'better' in some sense than traditional unbiased ones; several such instances will be discussed in this dissertation.

Stein-rule estimators also belong to the class of biased estimators. This general family of estimators is named for Charles Stein (1956), who showed that under total squared error loss the traditional maximum likelihood estimator of the mean of a multivariate normal random vector was inadmissible if the number of means to be estimated was greater than or equal to 3. Five years after this remarkable discovery, James and Stein (1961) were able to specify a simple nonlinear estimator which dominated the maximum likelihood estimator under squared error loss if 3 or more parameter restrictions exist and if certain other design related conditions are met. Stein's discovery has had a significant impact on the course of statistics over the past 25 years. In Scientific American, Bradley Efron

and Carl Morris (1979, p. 119) summarize the importance of Stein's contribution in the following way:

(Stein's) result undermined a century and a half of work on estimation theory, going back to Karl Fredrich Gauss and Adrien Marie Legendre. After a long period of resistance to Stein's ideas, punctuated by frequent and sometimes angry debate, the sense of paradox has diminished and Stein's ideas are being incorporated into applied and theoretical statistics.

In more recent formulations, Stein-rules have been proposed which incorporate uncertain prior information in the form of a testable (null) hypothesis in order to reduce the quadratic risk associated with having to estimate parameters of the linear statistical model. In practice, least squares parameter estimates are shrunk towards a hypothesized parameter vector. The degree of shrinkage is determined by the value of a random variable, usually the value of the statistic which tests the null hypothesis using the sample. Low numerical values of the test statistic indicate that the prior information is confirmed by the sample and suggests that the degree of shrinkage should be large. If the prior information is poor, then the test statistic is large; consequently, little if any shrinkage occurs and the Stein-rule is approximately equal to the ordinary least squares estimator.

It is well-known that estimator efficiency can be improved by imposing general linear restrictions on the parameters of the model; however, until Stein's breakthrough, it had not been shown that the level of bias induced by imposing inaccurate restrictions could be

controlled. Stein's rule for combining sample and nonsample information assures users that the bias-efficiency trade-off, as measured by mean square error, will on average be favorable when compared to the mean square error of the least squares estimator.

Before Stein-rule estimators become widely used, much more must be known about how they perform in practice. Not only do researchers want to know about the Stein-rule's sampling properties and robustness, they also would like to be shown the extent of possible improvements and how they may be applied to concrete problems. These issues are explored in the chapters that follow.

Surprisingly, little effort has been expended to determine how well a Stein-rule estimator might predict future values of the dependent variable of a linear equation. Although a superior fit within the sample is no guarantee that the future observations will be accurately predicted by a regression model, it is nevertheless reasonable to expect models characterized by 'improved' parameter estimates to yield better forecasts. One question worth considering is whether combining sample and nonsample information with a Stein-rule will lead to better out-of-sample forecasts than OLS, RLS, or ARIMA, especially when there is reason to believe that the process generating the in-sample data is different from the one generating the post-sample data.

Stein-rule estimators are often considered

impracticable because of the difficulty one has constructing confidence intervals and testing hypotheses. Conventional hypothesis testing and interval estimation procedures require knowledge of the estimator's covariance matrix as well as the exact or limiting distributions of certain linear and quadratic forms and their ratios. Many of the standard results of conventional theory are based on the normality or asymptotic normality of linear combinations of the estimators in question. Unfortunately, a Stein-rule estimator is nonlinear, nonnormally distributed, and its covariance matrix and other 'statistics' contain unknown population parameters. Replacing unknown population parameters with estimators yields statistics with unknown sampling properties and confidence intervals and ellipsoids cannot be obtained in the usual way. Chapter 5 of this dissertation is directed toward finding an operational way to address this problem.

Finally, preliminary research will be conducted examining the risk characteristics of a shrinkage estimator of the parameter of a nonlinear model. As a first step, the probit regression model will be considered. Nelder and Wedderburn (1972) have recently found the probit model to be imbedded in a class of generalized linear models; one consequence of this is that the parameters of the probit regression model can be estimated using iterated feasible generalized least squares. Within this framework, it is possible to consider a Stein-like estimator of the probit

regression model which uses the value of a likelihood ratio test statistic to control shrinkage toward the hypothesized vector. Using the results of GLIM, an estimator which resembles the usual Stein estimator for the linear regression model can be obtained.

In the next three sections each of these issues will be discussed in more detail. In the final section a brief outline will be presented.

1.2 Out-of-Sample Forecasting With Stein-Rule Estimators

A primary function of economists is to forecast or predict economic events. Though hypothesis testing is an indispensable tool for revising or rejecting economic propositions, any theory must ultimately be judged by its ability to explain historical episodes or, more importantly, to predict future occurrences. Few theories are deemed 'robust' if they fail to meet this second criterion. Hypothesis testing and point estimation, it may be argued, are merely the means used to improve the quality of prediction. Therefore, the utility of an estimation technique might rest solely on its ability to generate more accurate predictions given a particular economic model.

There are several reasons to expect estimators of the Stein-family to show promise as good predictors. As noted earlier, Stein-rules allow would-be forecasters to combine prior information with sample information in a desirable way; the risk of using poor nonsample information is guaranteed to be no greater than not using it at all.

Although least squares minimizes the mean square error of prediction (PMSE) in the class of linear unbiased estimators, it does not do so in the class of unbiased estimators. In fact, Copas (1983) shows that under certain conditions the Stein-rule predictor can anticipate the deterioration of the post-sample fit and give uniformly lower PMSE than least squares.

In a subsequent paper, Jones and Copas (1986) investigate the robustness of shrinkage predictors to departures from the assumption that the post-sample regressor matrix is similar to the in-sample regressor matrix. The work of Jones and Copas suggests that the Stein-rule estimator may be robust to model misspecification in a prediction context, especially if the misspecification arises as the result of changes in the linear relationships among the regressor variables over the post-sample period. In Chapter 4, this proposition is explored using the well-known St. Louis equation [Andersen and Jordan (1968)].

The St. Louis equation posits that nominal GNP is a linear function of current and past values of monetary policy (M1 or M2) and fiscal policy (candidate variables include government purchases, government expenditures, high-employment surplus, etc.). Economic theory is often silent on the form of the variables (log-levels, first differences, rates of change) and the length and shape of the distributed lags for each variable. The current

practice is to let the data determine variable form, lag length and shape [Seaks and Allen (1984); Batten and Thornton (1983), (1984)]. The time series used to estimate the St. Louis equation have recently undergone a significant revision [U.S. Department of Commerce, Bureau of Economic Analysis, (Dec. 1985)], and, not surprisingly, lead to a different empirical specification of the model. In addition, changes in Federal Reserve operating procedure in 1979 and again in 1982 lead many to believe that the underlying economic relationship between monetary policy and GNP had changed [Wallich (1984); Gilbert (1985)]. In light of the inherent uncertainty of model specification, the St. Louis equation is typical of many economic models and can be used to demonstrate the robustness of various prediction techniques to conditions normally encountered in econometric practice.

In Chapter 4, the post-sample forecast accuracy of several estimators of the St. Louis equation are compared using the root-mean-square error criterion. Several prediction equations will be estimated using the 1962:2-1982:3 and 1962:2-1979:3 sample periods for maximum lag lengths of 12 quarters. These include: 1) the OLS estimator; 2) various estimators suggested by model selection procedures; 3) a simple Stein-rule estimator which shrinks all coefficients toward zero [Judge and Bock, (1978)]; 4) a Stein-rule estimator which shrinks all slope coefficients toward zero [Lindley, (1961)]; 5) a Stein-rule

estimator which shrinks towards the principal components estimator [Hill and Judge, (1987)]; 6) a Stein-rule estimator which shrinks the parameter estimates toward the values implied by the estimators obtained through model selection; and, 7) A univariate ARIMA model [Box and Jenkins, (1976)]. Conditional forecasts will be generated from the estimated models for the 1979:4-1983:3 and 1982:4-1986:3 post-sample periods.

1.3 Confidence Intervals and Ellipsoids

If Stein-rule estimation is to ever gain widespread acceptance among applied economic researchers, an acceptable measure of precision must be developed. Although there are exceptions (ridge estimator, pretest estimator, etc.), knowledge of the sampling distribution of an estimator is an important pre-condition for its use. The exact covariance matrix of the Stein-rule estimator is known [Judge and Bock, Section 8.9, (1978)], but the formula contains unknown population parameters. As a consequence, interval estimates and hypothesis tests cannot be formulated in the usual fashion. If one attempts to replace the unknown parameters with estimates, the sampling distribution of the usual likelihood ratio is no longer Student's t or Snedecor's F .

Phillips (1984) has been able to show that the exact distributional properties are, as he puts it, "well within reach." Specifically, he provides a formula for the probability density of the James-Stein (1961) estimator of

the linear regression model and deduces moment formulae directly from this general result. Unfortunately, Phillip's results rely on the use of advanced analysis (Weyl fractional calculus) and cannot be readily applied at this time.

Less precise, but simpler, alternatives are available. Perhaps the easiest way to proceed is to approximate the sampling distribution of the statistics of interest by an asymptotic expansion. Ullah (1982) and Ullah, Carter, and Srivastava (1984) use an Edgeworth-type asymptotic expansion to approximate the multivariate and marginal sampling distributions for a class of biased estimators which includes those of the Stein-family and the corresponding overall F-statistic. Ohtani (1986) derives the distribution of an improved F-ratio [Ullah, Carter, and Srivastava, (1984)] obtained by using the James-Stein estimator in place of the OLS estimator and shows that the test based on the improved F-ratio for the null hypothesis that all regression coefficients are zero can be performed using the F-distribution. However, he also concludes that the power of this test is lower than that of the test given by the usual F-ratio.

Another line of research pursues Stein's (1962) conjecture that it is possible to derive improved confidence sets for the mean of a multivariate normal distribution. An improved confidence set is one with higher coverage probability and of no greater volume than

the usual one--a sphere or ellipsoid of fixed volume centered at the sample mean. Brown (1966) and Joshi (1967) independently demonstrated the existence of improved confidence sets when the multivariate normal random vector contains at least 3 elements. Olshen (1977) simulated the coverage probabilities of Joshi's estimator and found that the improvements could be substantial under certain parameterizations.

Using empirical-Bayes techniques, Morris (1977, 1983) shows that coverage probabilities of certain generalized-Bayes estimators are quite good. Using a Bayesian approach Berger (1980) constructs confidence ellipsoids based on the posterior covariance matrix. These ellipsoids are shown to have higher coverage probability over a significant portion of the parameter space and to be of uniformly smaller volume. Hwang and Casella (1982) devise an explicit procedure for uniformly increasing coverage probability by centering the usual confidence set at the positive-part James-Stein estimator. This result holds provided that the multivariate normal random vector has at least 4 elements. In addition, Hwang and Casella show that the possible improvement can be quite substantial.

The Berger (1980a) and Hwang and Casella (1982) estimators are limited to cases where the confidence sets are spherical. Hill and Fomby (1986) examine the coverage probability and volume of Berger's estimator relative to OLS under a range of conditions commonly found in

econometric practice. Surprisingly, they find Berger's estimator to be quite robust to various degrees of multicollinearity.

Most of the research on this topic has pursued the Bayesian confidence set approach. Many economists, however, are reluctant to embrace Bayesian statistics (one reason is that each new estimation problem requires a considerable start-up cost, thus being very time consuming). If progress is to be made, then an alternative must be found which not only yields tests of a given size with adequate power, but is also reasonably easy to perform.

One possibility is to use Efron's bootstrap [Efron (1979, 1981, 1987); Freedman (1981)] which is a general procedure for measuring the sampling variability of a statistic having an unknown sampling distribution. In essence, bootstrapping permits one to approximate the sampling distribution of a statistic by replacing the unknown distribution function with the empirical distribution of the data and then resampling randomly to obtain a Monte Carlo distribution of the resulting random variable. Chi and Judge (1985) have compared confidence intervals for the James-Stein estimator based on bootstrap resampling to those derived via empirical Bayes estimation under the assumption that σ^2 is known. The performance of bootstrap confidence ellipsoids has yet to be considered.

In Chapter 6, bootstrap confidence intervals and

ellipsoids are constructed using Efron's (1979) percentile method and the size and coverage probability of these are studied in a Monte Carlo experiment.

1.4 Risk Characteristics of a Stein-Like Estimator for the Probit Regression Model

Finally, a Stein-like shrinkage estimator of the probit regression model is considered. To date, the theory of Stein-rule estimation has not been extended to include nonlinear regression models. Despite the lack of an analytical result establishing a dominance property similar to the one for shrinking parameter estimates of a linear model, the idea that risk improvements measured under mean square error loss can be obtained in nonlinear models by shrinking maximum likelihood estimates toward hypothesized values is a reasonable one and warrants attention. In Chapter 6, a Stein-like shrinkage estimator for the probit regression model is proposed and its risk properties are studied in a Monte Carlo experiment.

The search for a Stein-like estimator of a nonlinear statistical model begins with the probit regression model for two reasons. First, the properties of the likelihood function for the probit model are well-known and understood; its probability density is said to be 'well-behaved' because it is regular and globally concave. Second, the MLE's can be interpreted as the result of an iterative generalized least squares procedure [Amemiya, (1985) and Nelder and Wedderburn, (1972)], a fact which immediately suggests an algorithm for constructing a Stein-

like shrinkage estimator. First, obtain the maximum likelihood parameter estimates for the probit model using iterated generalized least squares (or equivalent method). Then, calculate the value of the likelihood ratio statistic used to test the hypothesis restrictions. Finally, use this statistic to control the shrinkage of the unconstrained maximum likelihood estimates towards the restricted MLE's.

In this spirit, Dagenais (1985) extends ridge regression to nonlinear models by noting the weighted least squares interpretation of nonlinear least squares (NLLS) estimates and using these iteratively in the usual ridge regression procedure. Schafer, Roi, and Wolfe (1984) use a similar method to explore the statistical properties of a ridge logistic estimator. Copas (1983) has also suggested a shrinkage estimator for the probit model, but has not studied its risk properties.

In Chapter 6, the risk properties of a Stein-like shrinkage estimator for the probit regression model will be studied using Monte Carlo methods. In the experiment, risk functions under squared error loss will be computed and compared using the following estimators: maximum likelihood estimator (MLE), MLE over a restricted parameter space, a pretest estimator, and the proposed Stein-like estimator.

1.5 Plan

This dissertation consists of six additional chapters

and is organized in the following way. To make the volume self-contained, Chapters 2 and 3 will be devoted to summarizing existing theory and technical details used throughout the remainder of the work. In Chapter 2 the basic notions of statistical decision theory and biased estimation are introduced. It includes sections on 1) the use of nonsample information, 2) pretest estimation, 3) Stein-rule estimation, 4) hypothesis testing and interval estimation, and 5) the relationship between Stein-rules and Bayesian statistics. In Chapter 3 another set of traditional and nontraditional techniques are developed which are used in Chapters 4, 5, and 6. These include 1) time series models (ARIMA and PDL), 2) nonlinear estimation, 3) generalized linear models, and 4) computer intensive research techniques (Monte Carlo and bootstrapping).

In Chapter 4 the forecasting accuracy of the Stein-rule estimator of the St. Louis equation is considered. In Chapter 5 several means of deriving confidence intervals and ellipsoids using the bootstrap are considered and the size and coverage probabilities of these confidence procedures are explored in a Monte Carlo experiment. In Chapter 6 the risk characteristics of a Stein-like estimator of the probit regression model are compared to those of traditional estimators of the model, i.e., the MLE, restricted MLE, and pretest estimators. The final chapter will serve as a summary of results and a plan for future research.

Chapter 2

Biased Estimation

- 2.1 Classical Linear Regression Model
- 2.2 Statistical Decision Theory
 - 2.2.1 Decision Rules
 - 2.2.2 Loss Functions
 - (a) Squared Error Loss
 - (b) Linear Loss
 - (c) "0 - 1" Loss
 - (d) Risk Matrix
- 2.3 Alternative Estimation Rules
 - 2.3.1 Ordinary Least Squares
 - (a) Loss and Risk
 - (b) Normality
 - 2.3.2 Generalized Least Squares
 - 2.3.3 Bayesian Inference
 - (a) Prior Distribution Functions
 - (b) Point Estimation
 - 2.3.4 Empirical Bayes Inference
 - 2.3.5 James-Stein Estimator
- 2.4 The Hypothesis Restricted Regression Model
 - 2.4.1 Statistical Model, Mean, and Covariance
 - 2.4.2 Risk Under Weighted Quadratic Loss
 - 2.4.3 Risk Matrix
 - 2.4.4 Hypothesis Testing
- 2.5 Pretest Estimators
 - 2.5.1 Mean and Covariance
 - 2.5.2 Risk Under Weighted Quadratic Loss
 - 2.5.3 Risk Matrix
 - 2.5.4 Summary of OLS, RLS, and Pretest Estimation
- 2.6 Stein-Rule Estimators
 - 2.6.1 Statistical Model
 - (a) Positive-Part Rule
 - (b) Mean and Covariance
 - 2.6.2 Risk Under Weighted Quadratic Loss
 - 2.6.3 Risk Matrix
- 2.7 Stein-Rule Problems and Alternatives
- 2.8 Confidence Sets and Hypothesis Testing

Chapter 2 Biased Estimation

Classical statisticians use sample information in the form of observable random variables in conjunction with estimation rules to make inferences about unknown population parameters. In classical sampling theory an estimator is evaluated by its performance in repeated experimental trials. In this framework the statistician devises and uses estimation procedures which in a long series of identical experiments lead to correct parameter estimates (unbiased) with the greatest degree of accuracy possible (efficient). The benchmark for an estimator's goodness is the minimum variance unbiased property (m.v.u.).

In statistical decision theory, on the other hand, one explicitly considers the consequences of making decisions based on statistical information. The user of this information must specify a functional relationship that represents the rewards or costs of using an estimation rule to describe the true, but unknown state of nature. In the following two sections, the classical linear statistical model will be introduced and then linked to the basic principles of decision theory. In section 2.3 several estimators of the classical regression model will be considered (OLS, GLS, and MLE), as well as two alternatives to the frequentist approach (Bayesian and empirical Bayesian inference).

In section 2.4, the hypothesis restricted regression model is taken up and in sections 2.5 and 2.6 pretest estimation and Stein-rule estimation are treated, respectively. Then, several problems in Stein-rule estimation are noted and two solutions discussed in section 2.7. In the final section of the chapter, a general theory of confidence sets is summarized and related to standard principles of hypothesis testing.

2.1 Classical Linear Regression Model

Consider the linear model

$$y = X\beta + e \quad e \sim (0, \sigma^2 \Omega) \quad (2.1.1)$$

where y is a $T \times 1$ vector of observable random variables, X is a known $T \times K$ nonstochastic design matrix of rank K , β is a $K \times 1$ vector of unknown parameters, e is a $T \times 1$ vector of unobservable random variables with zero mean ($E[e]=0$), and finite variance, σ^2 is unknown and Ω is a known positive definite matrix. The set of assumptions underlying this model should be carefully considered before discussing the estimation of its parameters.

The $T \times 1$ vector y represents a sample of size T of the random variable Y . The y_i ($i = 1, 2, \dots, T$) may or may not be statistically independent. If not, then it is assumed that the researcher knows the covariance of y up to a scalar multiple σ^2 which must be estimated. Another important assumption of the classical linear regression model (CLRM) is that (2.1.1) represents all nonsample information and the sample values of y represent all sample

information about the unknowns β and σ^2 .

Also, the classical linear regression model assumes that the X matrix represents the design of a specific experiment. It is composed of T observations of the K treatment variables which are assumed to be under the control of the experimenter. Although the vector y varies from experiment to experiment, X is fixed in repeated samples and therefore would be identical in an infinite number of experiments. The strong assumptions required of X are hardly ever met in practice. A more accurate and less stringent statement of the underlying probability model would be that X is not repeated identically and that the probability distribution of y is conditional on the sample values of X as well as the population parameters β .

For maximum likelihood estimation of β it is also assumed that the researcher knows the joint probability distribution of $y|\beta, X, \Omega, \sigma^2$. The likelihood function is formed by rewriting the joint p.d.f. as $L(\beta, \sigma^2 | X, y, \Omega)$ which is then interpreted as the probability of obtaining all experimental values from the given parameterization β and σ^2 . The principle of maximum likelihood is to choose estimates $\hat{\beta}$ and $\hat{\sigma}^2$ which maximize the probability of generating the sample from the given experiment.

Given the model (2.1.1), the classical statistician attempts to find rules which use the available sample information y to derive the 'best' estimates of the unknown population parameters β . The rule chosen is influenced by

the fact that the researcher is able to either replicate the experiment or choose the number of experimental trials. Under these circumstances, it is not unreasonable to limit the class of estimators to that which yields unbiased estimates. Given enough replications or observations, the estimates derived will, on average, converge to the values of the population parameters.

Once the choice is limited to unbiased estimators, a frequentist will select from the set of unbiased rules that which, on average, gives the most precise estimates. That is, given two estimators of β , say β^* and β^{**} , β^* is more precise than β^{**} if $\text{Cov}(\beta^{**}) - \text{Cov}(\beta^*) = \Delta$, where Δ is a positive semidefinite matrix. If no other unbiased estimator of β can be found which is more precise than β^* , then β^* is said to be m.v.u.

2.2 Statistical Decision Theory

The term statistical decision theory refers to the class of problems in which the statistician must gain information about certain parameter values in order to make a decision when the consequences of that decision depend upon the unknown parameters.

Suppose that an experiment is designed to reveal information about the true state of nature, β . A decision d from the set of all possible decisions \mathbf{D} will be chosen based on the outcome of the experiment. Since the observation of y has a bearing on the decision chosen, d must depend on y , i.e., $d(y)$. Although it need not be, the

decision rule $d(y)$ is often used synonymously with the estimator of β . So, if $d(y)=\beta^*$, then $\beta^* \in \mathbf{D}$.

Let \mathbf{R} be the set of all possible rewards r which might be received as a result of the decision d and the true state of nature β . Therefore if $d(y)=\beta^*$, then

$$r(\beta, d(y)) = r(\beta, \beta^*) \in \mathbf{R}.$$

Let the statistician's or policy maker's utility function be denoted by \mathbf{U} and let \mathbf{U} be a function of the reward r . It is conventional to use the negative of the utility function rather than utility as the quantity of interest; this number is defined as the loss to the decision maker of having to use an estimator β^* instead of the true parameter value β as a basis for making the decision. Hence, for each state of nature $\beta \in \mathbf{B}$ and each decision $d \in \mathbf{D}$, the loss $L(\beta, d)$ is defined to be

$$L(\beta, d(y)) = - \mathbf{U}[r(\beta, d(y))]$$

or,

$$L(\beta, \beta^*) = - \mathbf{U}[r(\beta, \beta^*)].$$

The elements of a decision problem are the parameter space \mathbf{B} which reflects all possible states of nature relative to the unknown parameters β , the set of possible decisions \mathbf{D} , and a loss function $L(\beta, \beta^*)$ defined for all $(\beta, \beta^*) \in (\mathbf{B} \times \mathbf{D})$.

For any decision function $d(y) \in \mathbf{D}$ and parameter vector $\beta \in \mathbf{B}$, the risk function $\rho(\beta, d(y))$ is defined to be the mean value of the loss function over the sample space. Risk is denoted as

$$\rho(\beta, d(y)) = E[L(\beta, d(y))] = \int_{\mathbf{y}} L(\beta, d(y)) f(y|\beta) dy \quad (2.1.2)$$

or, if $d(y) = \beta^*$, then

$$\rho(\beta, \beta^*) = E[L(\beta, \beta^*)] = \int_{\mathbf{y}} L(\beta, \beta^*) f(y|\beta) dy \quad (2.1.3)$$

where $f(y|\beta)$ is the joint probability density of y given the true state of nature β and $\int_{\mathbf{y}}$ is the multiple integral over all possible values of the random vector y .

2.2.1 Decision Rules

Given the loss function $L(\beta, \beta^*)$ and the associated risk function $\rho(\beta, \beta^*)$ the question arises: What criteria can be used to determine which decision or estimator is to be preferred? One possible candidate is the estimator which has uniformly lower risk than other competing rules. Let β^* and β^{**} be two estimators in \mathbf{D} . Comparing β^* to β^{**} on the basis of the risk function using this criterion, β^* is preferred to β^{**} if

$$\begin{aligned} \rho(\beta, \beta^*) &\leq \rho(\beta, \beta^{**}) && \text{for all } \beta \in \mathbf{B} \text{ and} \\ \rho(\beta, \beta^*) &< \rho(\beta, \beta^{**}) && \text{for at least one } \beta \in \mathbf{B}. \end{aligned}$$

If no other estimator in \mathbf{D} (or equivalently, no other decision $d(y)$ in \mathbf{D}) is uniformly better than β^* , then β^* is called an admissible estimator (or d^* is called an admissible decision rule). To be more precise, consider the following definitions presented in Judge and Bock (1978).

Definition 2.1.1 An estimator β^* is said to dominate an estimator β^{**} if, for all $\beta \in \mathbf{B}$, $\rho(\beta, \beta^*) \leq \rho(\beta, \beta^{**})$. If, in addition, $\rho(\beta, \beta^*) < \rho(\beta, \beta^{**})$ for at least one $\beta \in \mathbf{B}$, then β^* strictly dominates β^{**} . [Judge and Bock (1978), p. 13]

Consider two estimators of β , β^* and β^{**} . Definition 2.1.1 says that if for any value of the true parameter vector β the estimator β^* has risk no greater than another estimator β^{**} , then β^* dominates β^{**} . If in addition, one can find at least one point in the entire parameter space where β^* is less risky than β^{**} , then β^* strictly dominates β^{**} . Clearly, the admissibility of an estimator is closely related to the idea of dominance. To see the exact nature of this relationship, consider the following definition.

Definition 2.1.2 An estimator β^{**} is said to be inadmissible, if, for any estimator β^* such that $\rho(\beta, \beta^*) \leq \rho(\beta, \beta^{**})$, for all $\beta \in \mathbf{B}$, and for some $\beta \in \mathbf{B}$, $\rho(\beta, \beta^*) < \rho(\beta, \beta^{**})$. [Judge and Bock (1978), p. 14]

By this definition, an estimator is inadmissible if it is not strictly dominated by another estimator.

Finally, another desirable quality for an estimator to have is that of minimaxity.

Definition 2.1.3 An estimator β^* is said to be minimax within the class of estimators \mathbf{D} if β^* is in \mathbf{D} and

$$\sup_{\beta \in \mathbf{B}} \rho(\beta, \beta^*) \leq \sup_{\beta \in \mathbf{B}} \rho(\beta, \beta^{**})$$

for all $\beta^{**} \in \mathbf{D}$. [Judge and Bock (1978), p. 14]

Note: This criterion merely states that the least upper bound of the risk associated with β^* is less than or equal to the least upper bound of the risk of all other

estimators in **D**. Estimators are called minimax if they minimize the maximum risk. Universal minimaxity is difficult to establish; therefore, the minimax property is confined to comparing estimators of the same class.

Bayes' criterion has also been used as a criterion for choosing an estimator. This approach will be discussed in sections 2.3.3 and 2.3.4.

2.2.2 Loss Functions

Loss functions can take many forms. Ideally, the loss function should be derived from an underlying utility function and the rewards on which it depends. Utility is difficult to model formally, consequently analyses of decisions are usually carried out under certain standard loss functions which may or may not accurately reflect the true losses to the decision maker.

(a) Squared Error Loss

For an arbitrary estimator \tilde{b} of β , the loss function $L(\beta, \tilde{b}) = (\beta - \tilde{b})'(\beta - \tilde{b})$ is referred to as squared error loss. It is used for several reasons [Berger (1980b), pp. 54-55]. First, squared error loss is the variance of the ordinary least squares estimator (or other unbiased estimator) in univariate cases and the trace of the variance-covariance matrix when β is vector valued. This fact makes risk evaluation under squared error loss familiar to classical statisticians. Another reported advantage of squared error loss functions is the relative ease with which risk can be calculated. On the other hand,

squared error loss is probably not a good model of true loss because it is neither bounded nor concave, two properties which violate common sense notions about a decision makers underlying utility function [i.e., decision makers probably have decreasing absolute risk aversion and the potential loss they suffer is usually bounded above, see DeGroot (1970), Ch. 7 or Hey (1981), p. 150].

Squared error loss is a special case of quadratic loss. If $\beta = (\beta_1, \beta_2, \dots, \beta_K)'$ is a vector to be estimated using $\tilde{b} = (\tilde{b}_1, \tilde{b}_2, \dots, \tilde{b}_K)'$ and W is a $K \times K$ positive definite matrix, then

$$L(\beta, \tilde{b}; W) = (\beta - \tilde{b})' W (\beta - \tilde{b}). \quad (2.2.1)$$

If $W = I_K$, then (2.2.1) is equivalent to squared error loss. If $W = X'X$, then (2.2.1) is called mean square error of prediction loss. To see why this is so, let $W = X'X$ and (2.2.1) becomes

$$L(\beta, \tilde{b}; W) = (\beta - \tilde{b})' X'X (\beta - \tilde{b}) = (X\beta - X\tilde{b})' (X\beta - X\tilde{b}). \quad (2.2.2)$$

Using the fact that $E[y] = X\beta$ and by denoting the predicted values of y as \hat{y} , (2.2.2) can be written as

$$L(\beta, \tilde{b}; W) = (\hat{y} - E[y])' (\hat{y} - E[y]). \quad (2.2.3)$$

This loss function is used when interest focuses on the in-sample forecast accuracy of the model, whereas the squared error loss function is considered to be more appropriate when interest lies in the 'quality' of the parameter estimates.

(b) Linear Loss

When the utility function is approximately linear the loss function tends to be linear. In this instance the loss function of interest is

$$L(\beta, b) = \begin{cases} K_0(\beta - \tilde{b}) & \text{if } \beta - \tilde{b} \geq 0 \\ K_1(\beta - \tilde{b}) & \text{if } \beta - \tilde{b} < 0. \end{cases} \quad (2.2.4)$$

The constants K_0 and K_1 may be chosen so that overestimation is valued either more or less than underestimation, i.e., $K_0 \leq K_1$. If $K_0 = K_1 = K$, then the loss function (2.2.4) becomes

$$L(\beta, \tilde{b}) = K|\beta - \tilde{b}| \quad (2.2.5)$$

and is called absolute error loss. The constants K_0 and K_1 may also be functions of the true state of nature β , in which case the loss is called weighted linear loss.

[Berger (1980b), p. 56]

(c) "0 - 1" Loss

In many cases, a decision maker must select one of two alternatives. Hypothesis testing falls into this category of decision problems. Loss functions in this class are denoted

$$L(\beta, b) = \begin{cases} 0 & \text{if } \beta \in B_i \\ 1 & \text{if } \beta \in B_j \end{cases} \quad (i \neq j). \quad (2.2.6)$$

If the correct decision is made, then the loss is zero. If not, the loss is equal to one. The risk associated with this loss function is

$$\rho(\beta, \tilde{b}) = E[L(\beta, \tilde{b})] = \Pr[\tilde{b} \text{ is correct}] \quad (2.2.7)$$

which can be interpreted as either the probability of a

type I or type II error, depending on whether $\beta \in \mathbf{B}_i$ or $\beta \in \mathbf{B}_j$ (Under $H_0: \beta \in \mathbf{B}_i$ and under $H_a: \beta \in \mathbf{B}_j$).

(d) Risk Matrix

Another related measure of estimator performance, the risk matrix, is often used. The mean square error matrix of an estimator b is defined to be

$$E[(\beta - \tilde{b})(\beta - \tilde{b})'] \equiv \text{MSE}(\beta, \tilde{b}).$$

That is,

$$\begin{aligned} \text{MSE}(\beta, \tilde{b}) &= E\{[E(\tilde{b}) - \beta][E(\tilde{b}) - \beta]'\} \\ &\quad + E\{[\tilde{b} - E(\tilde{b})][\tilde{b} - E(\tilde{b})]'\} \\ &= [\text{bias}(\tilde{b})][\text{bias}(\tilde{b})]' + \text{Cov}(\tilde{b}). \end{aligned}$$

Notice that under squared error loss

$$\text{tr}[\text{MSE}(\beta, \tilde{b})] = \rho(\beta, \tilde{b}).$$

If \tilde{b} and b^* are two alternative estimators of the vector β , then b^* is defined to be superior to \tilde{b} in strong mean square error if and only if

$$E[(\tilde{b} - \beta)(\tilde{b} - \beta)'] - E[(b^* - \beta)(b^* - \beta)'] = \Delta$$

where Δ is a positive semi-definite matrix. [Fomby, Hill, and Johnson (1984), p. 98]

2.3 Alternative Estimation Rules

In this section, several estimators of the linear model (2.1.1) will be considered. Estimators include the ordinary least squares estimator (OLS), the generalized least squares estimator (GLS), and the maximum likelihood estimator (MLE). In addition, Bayesian and empirical Bayesian approaches to point estimation are briefly considered.

2.3.1 Ordinary Least Squares

Suppose that one wishes to estimate the vector of unknown parameters β of the linear model (2.1.1). For the classical statistician the problem is to find a suitable function of the observed random variables y that yields the 'best' estimator of β in repeated samples. Under the assumption that $\Omega = I_T$, the most widely used rule is that of ordinary least squares (OLS). OLS chooses that value of β which minimizes the sum of squared errors function of the model (2.1.1), which is denoted

$$s \equiv e'e = (y - X\beta)'(y - X\beta).$$

Minimizing s with respect to β

$$\frac{\partial s}{\partial \beta} = 2X'Xb - 2X'y = 0$$

yields,

$$b = (X'X)^{-1}X'y \quad (2.3.1)$$

where b is defined to be the ordinary least squares (OLS) estimator of β . The OLS estimator of β is linear and unbiased. Linearity follows from the fact that $(X'X)^{-1}X'$ is a $K \times T$ matrix of constants and unbiasedness from the fact that $E[b] = \beta$. The covariance matrix of the OLS estimator is denoted

$$\text{Cov}(b) = E[(b - \beta)(b - \beta)'] = E[(X'X)^{-1}X'ee'X(X'X)^{-1}]$$

and, since $E[ee'] = \sigma^2 \Omega = \sigma^2 I_T$, this reduces to

$$\text{Cov}(b) = \sigma^2 (X'X)^{-1}. \quad (2.3.2)$$

In summary, $b \sim (\beta, \sigma^2 (X'X)^{-1})$. The unbiased estimator of the unknown parameter σ^2 is denoted

$$\hat{\sigma}^2 = (y - Xb)'(y - Xb)/(T - K). \quad (2.3.3)$$

By the Gauss-Markov theorem, the least squares estimator is the most efficient linear unbiased estimator of β for the model (2.1.1), given $\Omega = I_T$. [Rao (1976), pp. 223-224].

(a) Loss and Risk

The risk function of the OLS estimator under weighted quadratic loss is

$$E[(b - \beta)'W(b - \beta)] = \sigma^2 \text{tr}[(X'X)^{-1}W] \quad (2.3.4)$$

Notice that under mean square error loss (i.e., $W = I_K$), equation (2.3.4) is merely the trace of the covariance matrix of b . Judge and Bock (1978, pp. 19-20) show that within the class of linear unbiased estimators, the least squares estimator of β is also minimax.

(b) Normality

Under the additional assumption that e is distributed as a multivariate normal random vector, one can make further claims about the goodness of the linear estimator $b = (X'X)^{-1}X'y$. It can be shown that when $e \sim N(0, \sigma^2 I)$, b is the maximum likelihood estimator as well as the m.v.u. estimator of β . The maximum likelihood estimator of σ^2 is biased, however, and denoted $\hat{\sigma}^2 = \hat{e}'\hat{e}/T$. [see Fomby, Hill, and Johnson (1984), pp. 34-35]. Finally, it can be shown that b is minimax among all unbiased (linear or nonlinear) estimators of β [Judge and Bock (1978), p. 20].

2.3.2 Generalized Least Squares

The generalized least squares estimator is used to estimate β in (2.1.1) when $\Omega \neq I_T$. Recall that Ω is assumed to be a known positive definite matrix; therefore, there exists another $T \times T$ positive definite matrix P such that $\Omega = PP'$. Transform (2.1.1) by P^{-1} to obtain:

$$P^{-1}y = P^{-1}X\beta + P^{-1}e. \quad (2.3.5)$$

Let $P^{-1}y = y^*$, $P^{-1}X = X^*$, and $P^{-1}e = e^*$. Equation (2.3.4) can now be written as,

$$y^* = X^*\beta + e^*. \quad (2.3.6)$$

Equation (2.3.6) is sometimes referred to loosely as the transformed model. Observe that $E[e^*] = 0$ and $\text{Cov}(e^*) = E[P^{-1}ee'(P^{-1})'] = \sigma^2 P^{-1}\Omega(P^{-1})' = \sigma^2 I_T$. Thus, the transformed model (2.3.6) has the same properties as (2.1.1); consequently, the OLS estimator of the transformed model will retain all of its desirable properties.

The generalized least squares estimator of β may be written as

$$b_g = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y = (X^{*'}X^*)^{-1}X^{*'}y^* \quad (2.3.7)$$

and is, by the Gauss-Markov result, the best linear unbiased estimator of β . The GLS estimator of β has covariance matrix

$$E\{[b_g - E(b_g)][b_g - E(b_g)]'\} = \sigma^2 (X'\Omega^{-1}X)^{-1} \quad (2.3.8)$$

and the unbiased estimator of σ^2 is

$$\sigma_g^2 = (y - Xb_g)'\Omega^{-1}(y - Xb_g)/(T-K). \quad (2.3.9)$$

Given that Ω is known, the risk results from the previous

section hold for the transformed model (2.3.6). In summary,

$$b_g \sim [\beta, \sigma^2 (X' \Omega^{-1} X)^{-1}].$$

When Ω is unknown it can, in most instances, be replaced by a consistent estimator $\hat{\Omega}$ and the resulting feasible generalized least squares estimator (FGLS) will have the same asymptotic distribution as the GLS estimator [see Schmidt (1976), pp. 70-71].

2.3.3 Bayesian Inference

There is another important tradition in statistics which attacks the problem of estimating β in (2.1.1) in a different way. Philosophically, classical statisticians and Bayesians are often at odds because they disagree on certain fundamental principles (for instance, how to define probability). With the renewed interest among classical statisticians of using prior information efficiently, the two camps appear to be moving closer to one another (in spite of the continuing dispute over first principles). Nowhere is this more apparent than in Stein-rule estimation. The Stein-rule can be thought of as the classical statistician's response to the Bayesian's use of a prior probability density function and, remarkably enough, it can be derived though the empirical Bayesian approach as the mean of a posterior distribution [Zellner and Vandaele (1974)]

Given the close relationship between the Stein-rule and Bayesian approaches to point estimation, no thorough

examination of Stein estimation would be complete without a cursory discussion of the relationship between the two. Therefore, in the next 3 subsections these issues are examined. In the remainder of this section, the basic principles of Bayes estimation are presented and the relationship between the Bayes and the frequentist approach is discussed [Zellner (1971)]. In section 2.3.4, the empirical Bayes estimator is defined. Section 2.3.5 contains a sketch of how the simple James-Stein estimator (1961) is derived as an empirical Bayes estimator.

The decision theoretic framework developed in section 2.2 needs but slight alteration to accomodate Bayesians. Suppose the Bayesian statistician experiments to obtain information about the parameters β . The sample observations y depend on β through a known probability density $f(y|\beta)$. Let $f(y, \beta)$ denote the joint probability density function of the random vector of experimental observations y and the random parameter vector β . The parameter vector may have as its elements the coefficients of the linear regression model (2.1.1), its unknown variance σ^2 , or if Ω unknown, its covariances.

Using familiar properties of conditional probabilities, the joint density $f(y, \beta) = f(y|\beta)f(\beta)$ is obtained. From this, the marginal density $k(y)$ is obtained by integrating β out of the joint density. Then, one uses Bayes formula to find the distribution of interest, $f(\beta|y)$. Thus,

$$f(\beta|y) \propto f(y|\beta) f(\beta) \quad (2.3.10)$$

where \propto means "is proportional to", $f(\beta|y)$ is the posterior p.d.f. for the parameter vector β given the sample y , $f(\beta)$ is the prior p.d.f. for the parameters β , and $f(y|\beta)$ is the likelihood function and represents the sample information.

(a) Prior Distribution Functions

Specification of the prior probability density function can be problematic. Many Bayesians choose prior p.d.f.'s which when combined with the likelihood function yield posterior distributions that are easy to work with.

Definition 2.3.1 Let \mathcal{F} denote the class of probability density functions $f(y|\beta)$. A class of prior distributions \mathbf{P} is said to be a conjugate family for \mathcal{F} if the posterior p.d.f. $f(\beta|y)$ is in the class \mathbf{P} for all $f \in \mathcal{F}$ and prior probability density functions $\pi \in \mathbf{P}$.

When a conjugate prior is used, the resulting posterior p.d.f. can be reused as a prior p.d.f. in a subsequent experiment. In this way, researchers can update the posterior p.d.f. whenever additional sample information becomes available.

Another type of prior distribution often used is the improper prior distribution. If the integral of the prior distribution taken over the entire parameter space does not converge, then the prior distribution is called improper. This type of prior distribution poses no difficulty as long as the resulting posterior p.d.f. is proper. For example,

Jeffery's noninformative prior is typical of an improper prior density function and is denoted

$$f(\beta, \sigma) \propto 1/\sigma \quad -\infty < \beta < \infty, \quad 0 < \sigma < \infty.$$

Jeffery's prior p.d.f. has the attractive feature of being invariant under reparameterization. Noninformative priors are so-named because treat all possible values of β as equally likely.

(b) Point Estimation

The major object of Bayesian analysis is derivation of the posterior p.d.f. As mentioned above, this requires combining prior information (in the form of the prior p.d.f.) with the sample information (using the likelihood function). Once obtained, the posterior p.d.f. can be used to derive point estimates, find confidence intervals, or test hypotheses.

Again, following Zellner (1971), the relationship between the Bayesian and sampling theoretic approaches to point estimation can be depicted in the following way. Let $\tilde{b} = \tilde{b}(y)$ be the sampling theory estimate of β . Recall from equation (2.1.3) the risk function

$$\rho(\beta, \tilde{b}) = \int_y L(\beta, \tilde{b}) f(y|\beta) dy \quad (2.3.11)$$

where $L(\beta, \tilde{b})$ is the loss function, $f(y|\beta)$ is the p.d.f. of the sample given β , and the integral is taken over all values of y and is assumed to converge. Note, however, that the risk (2.1.11) is a function of the unknown parameter vector β , which the Bayesian considers to be a random variable. Consequently, we can consider the average

risk taken over all values of β ,

$$E[\rho(\beta, \tilde{b})] = \int_{\beta} f(\beta) \rho(\beta, \tilde{b}) d\beta. \quad (2.3.12)$$

In (2.3.12) $f(\beta)$ is the prior p.d.f. and is included in order to weigh the performance of the estimator \tilde{b} in various regions of the parameter space. This weighting is desirable because many estimators can be expected to perform "better" or "worse" depending on which region of the parameter space they operate.

To derive the point estimator, one may choose the value \tilde{b} which minimizes average risk; thus substitute (2.2.11) into (2.3.12) and select

$$\min_{\tilde{b}} E[\rho(\beta, \tilde{b})] = \min_{\tilde{b}} \int_{\beta} \int_Y f(\beta) L(\beta, \tilde{b}) f(y|\beta) dy d\beta. \quad (2.3.13)$$

Rearranging (2.3.13) yields

$$\min_{\tilde{b}} E[\rho(\beta, \tilde{b})] = \min_{\tilde{b}} \int_{\beta} [\int_Y L(\beta, \tilde{b}) f(y|\beta) dy] f(\beta) d\beta \quad (2.3.14)$$

or simply,

$$= \min_{\tilde{b}} \int_{\beta} \rho(\beta, \tilde{b}) d\beta. \quad (2.3.15)$$

By changing the order of integration in (2.3.14) and using the fact that $f(\beta)f(y|\beta) = f(y)f(\beta|y)$, (2.3.14) can be expressed as

$$\min_{\tilde{b}} E[\rho(\beta, \tilde{b})] = \min_{\tilde{b}} \int_Y [\int_{\beta} L(\beta, \tilde{b}) f(\beta|y) d\beta] f(y) dy. \quad (2.3.16)$$

The estimator \tilde{b} which minimizes the expression in brackets in (2.3.16) minimizes expected risk and this estimator is, by definition, the Bayes estimator.

A formal definition from Judge and Bock (1978) uses

(2.3.15) and is stated below.

Definition 2.3.2 An estimator β_f is said to be Bayes with respect to the distribution F in \mathbf{B} if the Bayes Risk (the expected value of the risk function with respect to the distribution F on \mathbf{B}) is minimum. That is, for all other β^* in \mathbf{B} ,

$$E[\rho(\beta, \beta_f^*)] = \int \rho(\beta, \beta_f^*) f(\beta) d\beta \leq E[\rho(\beta, \beta^*)]$$

[Judge and Bock (1978), p. 15]

Therefore, according to Definition 2.3.2, the Bayes estimator is that which minimizes the average risk with respect to the prior distribution F . If, for example, the loss function is given by $(\beta - b)'(\beta - b)$ [i.e., squared error loss], then the Bayes solution is the mean of the posterior distribution. Finally, the estimator b_{gb} which minimizes (2.3.16) for the posterior p.d.f. (2.3.10) when $\int f(\beta) d\beta$ is infinite or improper is referred to as a generalized Bayes estimator.

2.3.4 Empirical Bayes Inference

In the preceding section the Bayesian approach to estimation and decision theory was introduced. In this section, empirical Bayes estimators are considered; these estimators share features of the Bayesian and the classical sampling theory approaches to statistical inference. The distinguishing feature of empirical Bayes inference is the estimation of the parameters of the prior probability density function. Following Berger (1980b), the empirical

Bayes approach to statistical inference will be described in this section.

As in pure Bayesian analysis, the data are assumed to be distributed according to a particular family of distributions $f(y|\beta)$. The parameters β are themselves random variables with distribution $f(\beta) \in \mathbf{P}$ defined on the sample space \mathbf{B} . Inferences are to be made about a particular realization of β . Therefore,

$$y|\beta \sim f(y|\beta) \text{ and } \beta \sim f(\beta), f \in \mathbf{P} \quad (2.3.17)$$

where $f \in \mathbf{P}$ denotes the fact that $f(\beta)$ belongs to a family of prior density functions. Note also that (2.3.17) implies that there are two random processes to consider: one for the data y and one for the parameters β .

Again, given a loss function $L(\beta, \tilde{b})$, the associated empirical Bayes risk function is defined to be

$$\rho(\beta, \tilde{b}) = E_f E_\beta L(\beta, \tilde{b}), \quad f \in \mathbf{P}. \quad (2.3.18)$$

Notice that the expectation is taken over β and over all priors f so that the estimator (or decision procedure) is evaluated with respect to both sources of possible variability.

The model (2.3.18) is important because it contains both pure Bayesian and frequentist models as special cases. Pure Bayesians consider the case where the family of priors is restricted to a single member f_0 . Given f_0 and quadratic loss, application of Bayes rule yields the mean of the posterior distribution as a point estimate of β . On the other hand, if \mathbf{P} contains all point priors $f_\beta, \beta \in \mathbf{B}$, then

$\rho(\beta, \hat{b}) = \rho(f_\beta, b)$ becomes the ordinary risk function of the OLS estimator of β , $\rho(\beta, b)$.

2.3.5 James-Stein Estimator

It is possible to show that the James-Stein (1961) estimator of the mean of a multivariate normal random vector can be derived via the empirical Bayesian approach to statistical inference. The first step in deriving the empirical Bayes estimator is to transform the model into its canonical form. Let P be an orthogonal, positive definite matrix such that $P'X'XP=D$ with $D=\text{diag}[\lambda_1, \lambda_2, \dots, \lambda_K]$, and $\lambda_1 \geq \lambda_2 \geq \dots \lambda_K$, where λ_i ($i=1, \dots, K$) is the i^{th} characteristic root of the regressor cross product matrix $X'X$. Using the fact that $P'P=PP'=I_K$, define $Z=XP$ and $\theta=P'\beta$. Under this parameterization the model can be written

$$y = X\beta + e = XPP'\beta + e = Z\theta + e.$$

The least squares estimator of θ is denoted $\hat{\theta}$, $\hat{\theta} = \theta + \epsilon$, $\hat{\theta} \sim N(\theta, \sigma^2 D^{-1})$ and consequently, $\epsilon \sim N(0, \sigma^2 D^{-1})$.

The overall strategy is to find the posterior distribution of θ . This requires knowledge of the joint density $(\hat{\theta}, \theta)$ from which the conditional density $\theta | \hat{\theta}$ can be obtained. The latter is the posterior probability distribution of the parameters. Under quadratic loss, the mean of the posterior distribution is the Bayes estimator. Since the parameters of the prior distribution are estimated using the sample, one is said to be using an empirical Bayes estimator of θ .

The James-Stein estimator shrinks least squares coefficient estimates toward zero. In Bayesian analysis this is equivalent to invoking a prior distribution on θ which has 0 mean and an as yet unspecified covariance matrix. In the absence of any specific prior information about the covariance of θ , it is assumed that it is proportional to the sample covariance matrix of $\hat{\theta}$, i.e., $\theta \sim N(0, \tau^2 D^{-1})$, $\theta = 0 + v$, and $v \sim N(\tau^2 D^{-1})$.

To obtain the joint distribution $(\hat{\theta}, \theta)$, the marginal distribution $m(\hat{\theta})$ is needed. By substitution, $\hat{\theta} = 0 + v + \epsilon$ and assuming v and ϵ are statistically independent, $\hat{\theta}$ has covariance

$$E[(v + \epsilon)(v + \epsilon)'] = E[vv' + \epsilon\epsilon'] = (\tau^2 + \sigma^2)D^{-1} \quad (2.3.17)$$

and thus $\hat{\theta} \sim N(0, (\tau^2 + \sigma^2)D^{-1})$. The covariance $\text{Cov}(\hat{\theta}, \theta)$ is

$$E[(v + \epsilon)v'] = \tau^2 D^{-1} \quad (2.3.18)$$

Using the properties of the multivariate normal p.d.f.

[Dhrymes (1974), p. 19], the joint distribution $(\hat{\theta}, \theta)$

becomes

$$\begin{bmatrix} \hat{\theta} \\ \theta \end{bmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} (\tau^2 + \sigma^2)D^{-1} & \tau^2 D^{-1} \\ \tau^2 D^{-1} & \tau^2 D^{-1} \end{pmatrix} \right] \quad (2.3.19)$$

and the conditional distribution $\theta | \hat{\theta}$ is

$$\theta | \hat{\theta} \sim N[0 + (\tau^2 / (\tau^2 + \sigma^2))\hat{\theta}, \sigma^2 (\tau^2 / (\tau^2 + \sigma^2))D^{-1}]. \quad (2.3.20)$$

The Bayes estimator under quadratic loss is the mean of the posterior p.d.f.

$$\begin{aligned} E[\theta | \hat{\theta}] &= 0 + [\tau^2 / (\tau^2 + \sigma^2)]\hat{\theta} \\ &= [1 - \sigma^2 / (\tau^2 + \sigma^2)]\hat{\theta}. \end{aligned} \quad (2.3.21)$$

Since σ^2 and $(\tau^2 + \sigma^2)^{-1}$ are unknown, they must be estimated.

Note that since $\hat{\theta} \sim N(0, (\tau^2 + \sigma^2)D^{-1})$, then $\hat{\theta}'D\hat{\theta} \sim \chi_K^2(\tau^2 + \sigma^2)$ and thus

$$\begin{aligned} E[1/\hat{\theta}'D\hat{\theta}] &= (\tau^2 + \sigma^2)^{-1} E[1/\chi_K^2] \\ &= (\tau^2 + \sigma^2)^{-1}/(K-2) \end{aligned}$$

implying,

$$(\tau^2 + \sigma^2)^{-1} = (K-2)/\hat{\theta}'D\hat{\theta}. \quad (2.3.22)$$

Under quadratic loss, the risk minimizing scale invariant estimator of σ^2 is $\tilde{\sigma}^2 = (y - Z\hat{\theta})'(y - Z\hat{\theta})/T - K + 2 = s/T - K + 2$ which if used in (2.3.21) yields

$$\hat{\theta}_{eb} = [1 - (s/T - K + 2)(K-2)/\hat{\theta}'D\hat{\theta}]\hat{\theta}. \quad (2.3.23)$$

Applying the inverse transformation $\hat{\theta} = Pb$ yields the James-Stein estimator

$$b_{eb} = \delta(b) = [1 - (s/T - K + 2)(K-2)/b'X'Xb]b \quad (2.3.24)$$

where the shrinkage constant is chosen to be equal $(K-2)/(T-K+2)$. The estimated covariance matrix can be obtained similarly by noting

$$\sigma^2(\tau^2/\tau^2 + \sigma^2)D^{-1} = \sigma^2(1 - \sigma^2/\tau^2 + \sigma^2)D^{-1}.$$

Replacing σ^2 with $s/T - K + 2$ and $(1 - \sigma^2/\tau^2 + \sigma^2)$ with $[1 - (s/T - K + 2)(K-2)/b'X'Xb]$ yields

$$\text{Cov}(b_{eb}) = (s/T - K + 2)[1 - (s/T - K + 2)(K-2)/b'X'Xb](X'X)^{-1}. \quad (2.3.25)$$

This estimator not useful in some circumstances. Note that while $0 \leq (1 - \sigma^2/\tau^2 + \sigma^2) \leq 1$, its estimator (2.3.25) may be less than zero if $v = (s/T - K + 2)(K-2)/b'X'Xb > 1$. This leads to a nonpositive definite covariance matrix and standard errors which may be complex. In such cases, it is advisable to set v to its theoretical lower bound, zero.

Berger and Berliner (1984) cite several advantages and disadvantages of approaching Stein-rule estimation this way. To its credit, formal Bayes estimation ensures that the prior information will be used correctly, increases the likelihood that the resulting estimator is admissible, and provides a framework for obtaining confidence sets based on the posterior p.d.f. The major disadvantages are: (1) the estimator must sometimes be expressed as numerical integrals and (2) the frequentist risk properties are often hard to verify. In addition, the variability associated with estimation of the parameters of the prior distribution is taken for granted in the derivation of the posterior.

Estimation of the parameters of the prior probability density function can usually prevent profound misspecification. Consequently, the empirical Bayes estimator often dominates its purely Bayesian alternatives.

2.4 The Hypothesis Restricted Regression Model

If the researcher has precise knowledge about hypothesized values of the parameters of the linear model (2.1.1) and imposes restrictions on the parameter space of the model, then estimator efficiency may be considerably enhanced. The resulting estimator is referred to as the hypothesis restricted estimator.

The hypothesis restricted estimator is a basic component of the Stein-rule used below in section 2.6. In fact, the Stein-rule estimator is actually a convex combination of the OLS and hypothesis restricted

estimators.

More generally, nearly any linear estimator can be thought of as a hypothesis restricted estimator. Implicitly, the linear model (2.1.1) reflects the imposition of an infinite number of restrictions; i.e., one for each of an infinite number of possible regressors which could have been included in the matrix X . For example, the polynomial distributed lag estimator considered in section 3.1.4 can be thought of as a hypothesis restricted estimator; in this case the researcher hypothesizes that the effects of lagged independent variables fall along a polynomial of a given degree and is able to express these hypotheses as a set of linear homogeneous equations.

For reasons which will become apparent, the hypothesis restricted estimator considered below will often be referred to as the restricted least squares (RLS) estimator; and, because of the central position this estimator takes in the work which follows, the RLS estimator will be discussed at some length.

2.4.1 Statistical Model, Mean, and Covariance

Recall the linear regression model (2.1.1),

$$y = X\beta + e \quad e \sim (0, \sigma^2 I_T) \quad (2.1.1)$$

and assume further that additional information exists in the form of $J \leq K$ independent linear hypotheses involving the unknown parameters β . Mathematically, this is expressed as $R\beta = r$ where R is $J \times K$ matrix of rank J , $J \leq K$, and r is a $J \times 1$ vector of known constants. Also, define $\omega = R\beta - r$, so that ω

is a $J \times 1$ vector of parameters representing the degree of error in the hypotheses.

By convention the hypotheses are assumed to be true, implying $\omega=0$ [Judge and Bock (1978), pp. 26-27]. Unfortunately, the economist seldom has nontrivial prior information of the form $R\beta=r$ which is exactly true. The consequences of $\omega \neq 0$ will be explored presently. First, we consider the hypothesis restricted estimator or restricted least squares estimator of the model (2.1.1) subject to $R\beta-r=0$.

The restricted least squares (RLS) estimator is found by minimizing the sum of squared errors from (2.1.1) subject to the constraint $R\beta=r$, i.e.,

$$y = X\beta + e \quad \text{subject to} \quad R\beta=r. \quad (2.4.1)$$

To obtain the RLS estimator, form the Lagrangian function

$$L = (y - X\beta)'(y - X\beta) + 2\lambda'(R\beta - r)$$

where λ is the $J \times 1$ vector of Lagrangian multipliers; then, maximizing L with respect to β and λ and rearranging yields the restricted least squares estimator b_r of β . The RLS estimator is denoted

$$b_r = b + S^{-1}R'[RS^{-1}R']^{-1}(Rb-r) \quad (2.4.2)$$

where $S=X'X$ and $b = S^{-1}X'y$ (i.e., b is the OLS estimator of β).

The RLS estimator of β has mean,

$$E[b_r] = \beta - S^{-1}R'[RS^{-1}R']^{-1}\omega \quad (2.4.3)$$

and covariance matrix,

$$\text{Cov}(b_r) = \Sigma_r = \sigma^2[S^{-1} - S^{-1}R'[RS^{-1}R']^{-1}RS^{-1}] \quad (2.4.4)$$

$$= \sigma^2(S^{-1} - C)$$

where $C = S^{-1}R'[RS^{-1}R']^{-1}RS^{-1}$. The restricted least squares estimator b_r is unbiased if and only if the linear hypotheses $R\beta = r$ are exactly true. Note also that $\text{Cov}(b) - \text{Cov}(b_r) = \sigma^2 C$, a positive semi-definite matrix. Therefore, the restricted least squares estimator is more efficient than the ordinary least squares estimator regardless of the degree of hypothesis error, ω . Even though the RLS estimator is always more efficient than the OLS estimator, it may be biased, suggesting that b_r may or may not be riskier than b under quadratic loss.

2.4.2 Risk Under Weighted Quadratic Loss

The risk of the RLS estimator under weighted quadratic loss is denoted

$$\begin{aligned} \rho(\beta, b_r; W) &= E[(b_r - \beta)' W (b_r - \beta)] \\ &= \text{tr}[\text{Cov}(b_r) W] \\ &\quad + \omega' (RS^{-1}R')^{-1} RS^{-1} W S^{-1} R' (RS^{-1}R')^{-1} \omega \end{aligned} \quad (2.4.5)$$

Notice that as the degree of hypothesis error ω increases, the risk of using the RLS estimator to estimate β increases. Expanding (2.4.5) yields

$$\begin{aligned} \rho(\beta, b_r; W) &= \sigma^2 \text{tr}\{[S^{-1} - S^{-1}R'(RS^{-1}R')^{-1}RS^{-1}]W\} \\ &\quad + \omega' (RS^{-1}R')^{-1} RS^{-1} W S^{-1} R' (RS^{-1}R')^{-1} \omega \end{aligned} \quad (2.4.6)$$

or,

$$\rho(\beta, b_r; W) = \rho(\beta, b; W) - \sigma^2 \text{tr} C W + 2\sigma^2 \gamma \quad (2.4.7)$$

where $C = S^{-1}R'[RS^{-1}R']^{-1}RS^{-1}$,

$$\gamma = \omega' (RS^{-1}R')^{-1} RS^{-1} W S^{-1} R' (RS^{-1}R')^{-1} \omega / 2\sigma^2,$$

$S = X'X$, and W is any known positive definite weight matrix.

From (2.4.7) one can see that in order for $\rho(\beta, b_r)$ to be less than or equal to $\rho(\beta, b)$, the following condition must be met

$$0 \geq \rho(\beta, b_r; W) - \rho(\beta, b; W) = -\sigma^2 \text{tr} CW + 2\sigma^2 \gamma$$

or,

$$\frac{1}{2} \text{tr} CW \geq \gamma = \omega' (RS^{-1}R')^{-1} RS^{-1} WS^{-1} R' (RS^{-1}R')^{-1} \omega / 2\sigma^2. \quad (2.4.8)$$

Now note the following two facts:

$$(1) \quad (RS^{-1}R')^{-1} = (RS^{-1}R')^{-\frac{1}{2}} (RS^{-1}R')^{-\frac{1}{2}}$$

This follows since there exists an orthogonal matrix P such that $P[\text{diag}\{\xi_1, \xi_2, \dots, \xi_J\}]P' = (RS^{-1}R')^{-1}$ where ξ_i , $i=1, 2, \dots, J$ are the characteristic roots of $(RS^{-1}R')^{-1}$, thus $(RS^{-1}R')^{-\frac{1}{2}} = P[\text{diag}\{\xi_1^{\frac{1}{2}}, \dots, \xi_J^{\frac{1}{2}}\}]P'$.

$$(2) \quad RS^{-1} WS^{-1} R' \text{ is a } J \times J \text{ positive definite matrix.}$$

Using these facts and theorems on the extrema of quadratic forms [Rao (1973), p. 62] the unknown γ can be bounded above and below by two known values. Thus,

$$\xi_S \leq \frac{\omega' (RS^{-1}R')^{-\frac{1}{2}} U (RS^{-1}R')^{-\frac{1}{2}} \omega / 2\sigma^2}{\omega' (RS^{-1}R')^{-1} \omega / 2\sigma^2} \leq \xi_L \quad (2.4.9)$$

where $U = (RS^{-1}R')^{-\frac{1}{2}} RS^{-1} WS^{-1} R' (RS^{-1}R')^{-\frac{1}{2}}$, $\lambda = \omega' (RS^{-1}R')^{-1} \omega / 2\sigma^2$, and ξ_S and ξ_L are, respectively, the smallest and largest characteristic roots of U . In terms of the notation developed above this condition reduces to

$$\lambda \xi_S \leq \gamma \leq \lambda \xi_L. \quad (2.4.10)$$

Using (2.4.10), the relationship between the weighted risk

function of the OLS estimator and the RLS estimator is shown to be:

$$\begin{aligned} \rho(\beta, b; W) - \sigma^2 \text{tr} CW + 2\sigma^2 \xi_S \lambda &\leq \rho(\beta, \beta_r; W) \\ &\leq \rho(\beta, b; W) - \sigma^2 \text{tr} CW + 2\sigma^2 \xi_L \lambda. \end{aligned}$$

From this expression note the following points: (1) When the hypotheses are correct (i.e., $\lambda=0$), the risk of using the RLS estimator of β is 'pinched' between equivalent numbers, implying that the risk under either bound is the same. Furthermore, when $\lambda=0$, the risk of using b_r is less than that of using b to estimate β . (2) Given X and R , the risk of using the RLS estimator increases monotonically and is unbounded as specification error $\lambda \rightarrow \infty$. (3) Finally, RLS is more risky than OLS (i.e., $\rho(\beta, b) - \rho(\beta, b_r) \leq 0$) if the noncentrality parameter $\lambda \geq \text{tr} CW / 2\xi_S$ and is less risky than OLS if $\lambda < \text{tr} CW / 2\xi_L$.

2.4.3 Risk Matrix

The RLS estimator has risk matrix

$$\begin{aligned} E[(b_r - \beta)(b_r - \beta)'] &= \sigma^2 S^{-1} - \sigma^2 S^{-1} R' (R S^{-1} R')^{-1} R S^{-1} \\ &\quad + S^{-1} R' (R S^{-1} R')^{-1} \omega \omega' (R S^{-1} R')^{-1} R S^{-1} \end{aligned} \quad (2.4.11)$$

or,

$$\begin{aligned} \sigma^2 S^{-1} - S^{-\frac{1}{2}} Q' \begin{bmatrix} Q1' & 0 \\ 0 & 0 \end{bmatrix} \\ \begin{bmatrix} \sigma^2 & -2\lambda\sigma^2 & 0 & 0 \\ 0 & \sigma^2 I_{J-1} & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} Q1 & 0 \\ 0 & 0 \end{bmatrix} Q S^{-\frac{1}{2}} \end{aligned} \quad (2.4.12)$$

where Q is the orthogonal matrix such that

$$QS^{-\frac{1}{2}}R'(RS^{-1}R')^{-1}RS^{-\frac{1}{2}}Q' = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \quad (2.4.12.a)$$

$Q1$ is another orthogonal matrix such that

$$Q1 QS^{-\frac{1}{2}}R'(RS^{-1}R')^{-1}\omega\omega'(RS^{-1}R')^{-1}RS^{-\frac{1}{2}}Q' Q1' \\ = \begin{bmatrix} \omega'(RS^{-1}R')^{-1}\omega & 0 \\ 0 & 0 \end{bmatrix} \quad (2.4.12.b)$$

and $\lambda = \omega'(RS^{-1}R')^{-1}\omega/2\sigma^2$. From (2.4.12) it is evident that $MSE(b) - MSE(b_r) = \Delta$, where Δ is a positive semi-definite matrix, when $1 - 2\lambda \geq 0$ or $\lambda \leq 1/2$. Thus, for a "small" degree of misspecification, the RLS estimator is superior to the OLS estimator in strong mean square error. Wallace (1972) has proposed an easy means of testing the null hypothesis $H_0: \lambda \leq 1/2$ against the alternative $H_a: \lambda > 1/2$.

In summary, the restricted least squares estimator (4.1.1) can be used to increase the precision of estimating the parameter vector β (in the sense that $Cov(b) - Cov(b_r)$ yields a positive semi-definite matrix) even when the restrictions imposed are incorrect. When the restrictions are correct b_r is unbiased; when the restrictions are nearly correct, b_r has lower risk than b under weighted quadratic loss. Otherwise, $\rho(b, b_r; W)$ increases monotonically and without limit as hypothesis error increases.

Economic theory, even at its best, yields less than exact information about possible parameter restrictions; therefore in light of the preceding discussion, it is important to keep in mind the potential danger of using the

restricted least squares estimator when uncertain prior information of the form $R\beta=r$ is available.

2.4.4 Hypothesis Testing

In order to test the compatibility of the sample information with the linear hypotheses the likelihood ratio statistic $u = (Rb-r)'[RS^{-1}R']^{-1}(Rb-r)/J\hat{\sigma}^2$ can be used. If the null hypothesis $H_0: R\beta-r=\omega=0$ is true, then u is distributed as a central F random variable with J and $T-K$ degrees of freedom (i.e., $u \sim F_{J, T-K}$). Under the alternative hypothesis $H_a: R\beta-r \neq 0$, u is distributed as a noncentral F random variable with J and $T-K$ degrees of freedom and noncentrality parameter $\lambda = \omega'(RS^{-1}R')^{-1}\omega/2\sigma^2$ (i.e., $u \sim F_{J, T-K, \lambda}$).

2.5 Pretest Estimators

In the preceding section it was shown that using good nonsample information can improve an estimator's risk performance and that poor nonsample information can impair it. Researchers often check the quality of nonsample information using the sample with convenient test statistics having well-known sampling properties. The nonsample information is either adopted or abandoned based on the outcome of this preliminary test of significance. For instance, in section 2.4.4 the likelihood ratio principle was used to test the null hypothesis that a linear combination of the parameters $R\beta$ is equal to a known constant r , against the alternative hypothesis that $R\beta \neq r$ using the statistic:

$$u = (Rb - r)' [RS^{-1}R']^{-1} (Rb - r) / J\hat{\sigma}^2 \sim F_{J, T-K, \lambda}.$$

Now, let c be some predetermined critical value from the $F_{J, T-K, \lambda}$ distribution. Notice that the critical value c depends on the level of the test α and on the noncentrality parameter λ of the distribution of the random variable u . The researcher chooses a desired level of confidence $(1-\alpha)$, assumes the null hypothesis is true, and using this assumption, selects the value of c from the standard central F table. If $u \leq c$, then the null hypothesis cannot be rejected and one uses the restricted least squares estimator of β . If, on the other hand, $u > c$, the null hypothesis is rejected and the OLS estimator is used. The resulting estimator is called the pretest estimator because it is the result of a preliminary test of the hypothesized restrictions $R\beta - r = 0$. The pretest estimator is denoted

$$b_{PT} = I_{[0, c)}(u)b_r + I_{[c, \infty)}(u)b \quad (2.5.1)$$

where $I_{[0, c)}(u)$ and $I_{[c, \infty)}(u)$ are indicator functions which take the value 1 if u falls within the subscripted intervals and zero otherwise. Judge and Bock (1978) refer to (2.5.1) as a "testimator" because it is a function of the data, the hypotheses, and the size of the test.

2.5.1 Mean and Covariance

The pretest estimator (2.5.1) has mean

$$E[b_{PT}] = \beta - h_{\lambda}(2)S^{-1}R'(RS^{-1}R')^{-1}\omega \quad (2.5.2)$$

and covariance

$$\begin{aligned} \text{Cov}(b_{PT}) = \Sigma_{PT} = \text{Cov}(b) - \sigma^2 h_{\lambda}(2) S^{-1} R' (R S^{-1} R')^{-1} R S^{-1} \\ + \{2h_{\lambda}(2) - h_{\lambda}(4) - [h_{\lambda}(2)]^2\} \\ S^{-1} R' (R S^{-1} R')^{-1} \omega \omega' (R S^{-1} R')^{-1} R S^{-1} \end{aligned} \quad (2.5.3)$$

where $h_{\lambda}(1) \equiv \Pr[\chi^2_{(J+1, \lambda)} / \chi^2_{(T-K)} < cJ/(T-K)]$. Notice that if the hypotheses are true, (i.e., $\omega=0$), then

$$\text{Cov}(b_{PT}) - \text{Cov}(b) = \Delta_{PT-O},$$

where Δ_{PT-O} is a positive semi-definite matrix. That is to say, the pretest estimator is more efficient than the OLS estimator when the hypotheses are correct. Using the orthogonal matrices Q and $Q1$ defined in (2.4.12.a) and (2.4.12.b), one can express the matrix Δ_{PT-O} in a way which makes the necessary conditions for more precise estimation of β obvious. Note,

$$\begin{aligned} \Delta_{PT-O} = -\sigma^2 S^{-\frac{1}{2}} Q' \begin{bmatrix} Q1' & 0 \\ 0 & 0 \end{bmatrix} \times \\ \begin{bmatrix} 2\{2h_{\lambda}(2) - h_{\lambda}(4) - [h_{\lambda}(2)]^2\} \lambda + h_{\lambda}(2) & 0 & 0 \\ 0 & h_{\lambda}(2) I_{J-1} & 0 \\ 0 & 0 & 0 \end{bmatrix} \\ \times \begin{bmatrix} Q1 & 0 \\ 0 & 0 \end{bmatrix} Q S^{-\frac{1}{2}} \end{aligned} \quad (2.5.4)$$

Thus, a necessary condition for $\text{Cov}(b_{PT}) - \text{Cov}(b) = \Delta_{PT-O}$ to be a positive semi-definite matrix is

$$\lambda = \omega' (R S^{-1} R')^{-1} \omega / 2\sigma^2 \leq h_{\lambda}(2) / \{2h_{\lambda}(2) - h_{\lambda}(4) - [h_{\lambda}(2)]^2\}.$$

The specification error λ must be less than the expression on the right-hand side of the inequality which is a function of sample size T , the number of parameters to be estimated K , the number of restrictions J , and the size of

the test α .

The fact that $\text{Cov}(b_{PT}) \geq \text{Cov}(b_R)$ for all β can be seen by looking at the difference $\Sigma_{PT} - \Sigma_R$ when $\omega=0$, i.e., when pretest estimation is most precise. Given $\omega=0$, the resulting difference is

$$\begin{aligned}\Sigma_{PT} - \Sigma_R &= [\text{Cov}(b) - h_\lambda(2)S^{-1}R'(RS^{-1}R')^{-1}RS^{-1}] \\ &\quad - \text{Cov}(b) - \sigma^2 S^{-1}R'(RS^{-1}R')^{-1}RS^{-1} \\ &= \sigma^2 S^{-1}R'(RS^{-1}R')^{-1}RS^{-1}[1 - h_\lambda(2)]\end{aligned}$$

which is a positive semi-definite matrix, since $0 \leq h_\lambda(1) \leq 1$ for positive integers l .

In summary, the pretest estimator b_{PT} may be more or less precise than the OLS estimator of β , depending on the degree of specification error inherent in the hypotheses. Under no circumstance, however, will b_{PT} be more precise than the RLS estimator b_R . Keep in mind, however, that precision alone is a poor standard of comparison; quadratic risk measures are preferred because they weigh both the precision and the bias of an estimator.

2.5.2 Risk Under Weighted Quadratic Loss

The risk of the pretest estimator under weighted quadratic loss is defined to be

$$\rho(\beta, b_{PT}; W) = E[(b_{PT} - \beta)' W (b_{PT} - \beta)] \quad (2.5.5)$$

for any positive definite weight matrix W . Squared error loss is defined as a special case of (2.5.5), where $W = I_K$. Under weighted quadratic loss, the pretest estimator has risk

$$\begin{aligned} \rho(\beta, b_{PT}; W) &= \sigma^2 S^{-1} W - \sigma^2 \text{tr}[RS^{-1}WS^{-1}R'(RS^{-1}R')^{-1}]h_\lambda(2) \\ &\quad + \omega'(RS^{-1}R')^{-1}RS^{-1}WS^{-1}R'(RS^{-1}R')^{-1}\omega\{h_\lambda(4)-2h_\lambda(2)\} \end{aligned} \quad (2.5.6)$$

or,

$$\begin{aligned} \rho(\beta, b_{PT}; W) &= \rho(\beta, b; W) - \sigma^2 h_\lambda(2) \text{tr}(V) + 2\sigma^2 \gamma [h_\lambda(4) - 2h_\lambda(2)] \\ \text{where } V &\equiv RS^{-1}WS^{-1}R'(RS^{-1}R')^{-1} \text{ and} \\ \gamma &\equiv \omega'(RS^{-1}R')^{-1}RS^{-1}WS^{-1}R'(RS^{-1}R')^{-1}\omega/2\sigma^2. \end{aligned}$$

If $\rho(\beta, b_{PT}; W) \leq \rho(\beta, b; W)$, then

$$\sigma^2 h_\lambda(2) \text{tr}(V) - 2\sigma^2 \gamma [h_\lambda(4) - 2h_\lambda(2)] \geq 0.$$

Again, using theorems dealing with the extrema of certain quadratic forms [Rao (1973), pp. 60-67] upper and lower bounds can be placed on the risk difference between the pretest estimator and the OLS estimator. Denoting this difference as ρ^{Δ}_{P-O} , it follows that

$$\begin{aligned} \sigma^2 [\text{tr}(V)h_\lambda(2) + 2\lambda(h_\lambda(4) - 2h_\lambda(2))] \xi_S &\leq \rho^{\Delta}_{P-O} \\ &\leq \sigma^2 [\text{tr}(V)h_\lambda(2) + 2\lambda(h_\lambda(4) - 2h_\lambda(2))] \xi_L \end{aligned}$$

where ξ_S and ξ_L are the smallest and largest characteristic roots of V . The pretest estimator has lower risk than the OLS estimator if

$\lambda \geq \text{tr}(V)/\{2[2-h_\lambda(4)/h_\lambda(2)]\xi_S\}$ and the OLS estimator has lower risk if

$$\lambda \leq \text{tr}(V)/\{2[2-h_\lambda(4)/h_\lambda(2)]\xi_L\}.$$

If, by chance,

$$\begin{aligned} \text{tr}(V)/\{2[2-h_\lambda(4)/h_\lambda(2)]\xi_L\} &\leq \lambda \\ &\leq \text{tr}(V)/\{2[2-h_\lambda(4)/h_\lambda(2)]\xi_S\}, \end{aligned}$$

then one cannot be sure which estimator has lower risk for given λ . In order to alleviate this uncertainty one would have to know the degree of misspecification ω . Note that

$\lim_{\lambda \rightarrow \infty} h_{\lambda}(1) = 0$, which implies that as specification error increases, the probability of the noncentral F random variable being less than $cJ/(T-K)$ gets smaller. Consequently, the risk of the pretest estimator increases, reaches a maximum, and then falls. As it falls, the risk of b_r converges to that of the OLS estimator from above. As a practical matter then, pretesting can protect the researcher in cases where the nonsample information is very poor.

2.5.3 Risk Matrix

The pretest estimator has risk matrix

$$\begin{aligned} E[(b_{PT} - \beta)(b_{PT} - \beta)'] &= \sigma^2 S^{-1} \\ &+ S^{-1} R' (RS^{-1} R')^{-1} \omega \omega' (RS^{-1} R')^{-1} RS^{-1} [2h_{\lambda}(2) - h_{\lambda}(4)] \\ &- \sigma^2 h_{\lambda}(2) S^{-1} R' (RS^{-1} R')^{-1} RS^{-1} \end{aligned} \quad (2.5.7)$$

or,

$$\begin{aligned} E[(b_{PT} - \beta)(b_{PT} - \beta)'] &= \sigma^2 S^{-1} - \sigma^2 S^{-\frac{1}{2}} Q' \begin{bmatrix} Q1' & 0 \\ 0 & 0 \end{bmatrix} \times \\ &\begin{bmatrix} 2\{2h_{\lambda}(2) - h_{\lambda}(4) - [h_{\lambda}(2)]^2\} \lambda + h_{\lambda}(2) & 0 & 0 \\ 0 & h_{\lambda}(2) I_{J-1} & 0 \\ 0 & 0 & 0 \end{bmatrix} \\ &\times \begin{bmatrix} Q1 & 0 \\ 0 & 0 \end{bmatrix} Q S^{-\frac{1}{2}} \end{aligned} \quad (2.5.8)$$

where again Q and $Q1$ are defined as in (2.4.12.a) and (2.4.12.b). From (2.5.8) it is evident that the risk matrix of the pretest estimator will be smaller than that of the OLS estimator if

$$\sigma^2 h_\lambda(2) \text{tr}(V) - 2\sigma^2 \lambda [h_\lambda(4) - 2h_\lambda(2)] \geq 0.$$

This is equivalent to the condition that

$$\lambda \leq 1/\{2[2 - h_\lambda(4)/h_\lambda(2)]\}.$$

Using Theorem 1, Section B.3, in Judge and Bock (1978) it can be shown that if $T-K \geq 2$, then

$$\min\{1, w_0[1 + (T-K-2)/(J+4)]\} \geq h_\lambda(4)/h_\lambda(2) \geq w_0$$

where $w_0 \equiv c/\{[(T-K)/J] + c\}$. The dominance condition holds under the following, easily calculated, condition

$$\lambda \leq 1/4 \leq 1/[2(2-w_0)] \leq 1/2$$

If the hypothesis error ω is zero, then the risk matrix of the pretest estimator reduces to

$$\text{MSE}(b) = \sigma^2 h_\lambda(2) S^{-1} R' (R S^{-1} R')^{-1} R S^{-1}$$

which is always less than that of the OLS estimator and greater than that of the RLS estimator.

More generally, when $\omega \neq 0$, it can be shown that $E[(b_{PT} - \beta)(b_{PT} - \beta)'] - E[(b - \beta)(b - \beta)'] = \Delta$, a positive semi-definite matrix, when $\lambda \leq 1/2\{2 - [1 - (h_\lambda(4))/(1 - h_\lambda(2))]\}$; and, since $h_\lambda(4) < h_\lambda(2)$, then $1 - h_\lambda(4) > 1 - h_\lambda(2)$ and $\lambda \leq 1/2 < 1/2\{2 - [1 - (h_\lambda(4))/(1 - h_\lambda(2))]\}$ or simply $\lambda \leq 1/2$. As long as specification error is less than or equal to $1/2$ then the restricted least squares estimator is better in terms of matrix mean square error than the pretest estimator. For specification error λ greater than $1/2$, one is better off under this measure using the pretest estimator. Finally, as $\lambda \rightarrow \infty$, the risk matrix of the pretest estimator converges from above to the risk matrix of the OLS estimator.

2.5.4 Summary of OLS, RLS, and Pretest Estimation

In the preceding sections, several estimators of β for the model (2.1.1) were considered. The OLS estimator is the best linear unbiased estimator of β and is minimax within its class, but is inadmissible for $K > 2$. The MLE is a minimum variance unbiased estimator of β and is minimax within the class of unbiased estimators, but is also inadmissible for $K > 2$. The restricted least squares estimator has lower risk than the OLS estimator over a relatively small portion of the feasible parameter space and is neither admissible nor minimax within its class. In fact, the risk of the RLS estimator under quadratic loss is unbounded as hypothesis error increases.

Many researchers "peek" at the quality of the nonsample information by performing a preliminary test of significance of exact linear hypotheses; the hypotheses are jointly adopted or abandoned based on the outcome of a statistical test. Unfortunately, the resulting estimator has properties which differ significantly from either the OLS estimator or the RLS estimator (or their maximum likelihood counterparts). The pretest estimator is unbiased only in the unlikely event that the hypotheses imposed are exactly true. As a result, the pretest estimator is a function of hypothesis error, the size of the test, the number of restrictions imposed, and the available degrees of freedom ($T-K$).

Under certain conditions, the pretest estimator may be

better in terms of weighted quadratic risk than either the OLS or RLS estimator. However, the OLS estimator dominates the pretest estimator over all but a tiny portion of the parameter space. The best comment one can make in favor of the pretest estimator is that it limits risk in the face of profound ignorance since at some point the risk of pretesting reaches a maximum and then declines toward that of the OLS estimator.

In addition, Judge and Bock (1978) have investigated the risk characteristics of the autocorrelation pretest estimator and found the actual losses associated with its use to be quite small. Their results suggest that the risk properties of different pretest estimators may vary substantially and should be considered case by case.

In the next portion of the chapter, the discussion will shift toward another class of estimators which, like the pretest estimator, is a combination of the restricted least squares estimator and the ordinary least squares estimator. Unlike b_{PT} , estimators from this class dominate the OLS estimator β for the model (2.1.1) when certain design related conditions are met.

2.6 Stein-Rule Estimators

In this portion of the chapter, the discussion shifts to a another member of the class of biased estimators: the Stein-rule. The particular Stein-rule estimator considered here is in many respects similar to a pretest estimator. However, instead of accepting or rejecting the hypotheses based on a preliminary test of significance, the Stein-rule estimator is formed by taking a convex combination of the RLS and OLS estimators; in effect, least squares parameter estimates are "shrunk" toward the restricted least squares estimates by a degree determined by the quality of the restrictions imposed. The quality of the restrictions is gauged by the value of the usual F-statistic used to test general linear hypotheses. Low numerical values of the test statistic indicate that the restrictions are supported by the sample and that the degree of shrinkage should be large. If the data do not support the restrictions, little or no shrinkage occurs and the Stein-rule is approximately equal to OLS. More importantly, the Stein-rule estimator dominates the MLE under certain design related conditions and is itself dominated by a rather simple modification. In this section, a very general version of the Stein-rule estimator will be presented and the conditions for its dominance over the maximum likelihood (OLS) estimator will be examined.

2.6.1 Statistical Model

Recall the linear regression model

$$y = X\beta + e \quad e \sim N(0, \sigma^2 I_T). \quad (2.6.1)$$

This is the model (2.1.1) under the additional assumption that the random disturbances are normally distributed. Consequently, for (2.6.1) the MLE and OLS estimators of β are identical to one another and can be used interchangeably. Note also, the RLS estimator b_r has a similar interpretation if the log of the normal likelihood function is maximized with respect to β and σ^2 subject to the J independent restrictions $R\beta - r = 0$. Using these facts, the Stein-rule estimator for $J > 2$ is given below:

$$\delta = [1 - a(T-K)/Ju](b - b_r) + b_r \quad (2.6.2)$$

or,

$$\delta = [1 - a(T-K)/Ju]b + [a(T-K)/Ju]b_r \quad (2.6.3)$$

where $b = S^{-1}X'y$, $b_r = b + S^{-1}R'(RS^{-1}R')^{-1}(Rb - r)$,

$u = (Rb - r)'(RS^{-1}R')^{-1}(Rb - r)/J\hat{\sigma}^2$, $\hat{\sigma}^2 = (y - Xb)'(y - Xb)/(T-K)$,

and "a" is a non-negative shrinkage constant. The estimator (2.6.2) is mentioned in Judge and Bock [(1978), p. 241], but not developed and is a special case of an estimator proposed by Mittelhammer (1984).

It is worth noting at this point that the normality of the random error terms may not be a critical assumption in Stein estimation. Some work has been done in an attempt to extend Stein estimation to models characterized by nonnormal errors. Miyazaki, Judge, and Yancy (1986) and Judge, Miyazaki, and Yancy (1985) have explored linear

models with spherically symmetric errors (which include multivariate t, Cauchy, and normal) and conclude that the risk characteristics for traditional Stein-like estimators under normality are similar to those for the nonnormal cases.

The random variable u is the familiar likelihood ratio statistic for the test of the null hypothesis $R\beta=r$ against the alternative $R\beta \neq r$ and is distributed as a central $F_{J,T-K}$ if the null hypothesis is true. As the probability of u being from a $F_{J,T-K}$ distribution declines (i.e., as u increases), the nonsample information is weighted less heavily. In the limit, the weight given the OLS estimator is 1 and that given the RLS estimator is zero. If $u=a$, then the RLS estimator receives a weight of 1 and the OLS estimator a weight of zero. Unfortunately, u may be less than a , in which case, the sign of the OLS estimates is reversed. To many this represents a serious drawback of the Stein-rule estimator (2.6.2). In response, a so-called positive-part Stein-rule [Baranchik (1964)] has been proposed which sets $\delta=b_r$ whenever $u < a$. Berger and Bock (1975) prove that such an estimator dominates the usual Stein-rule under squared error loss.

(a) The Positive-Part Rule

The proposed positive-part Stein-rule is

$$\delta^+ = [1-a(T-K)/Ju](b-b_r) \times I_{[a,\infty)}[(b-b_r)'S(b-b_r)/s](b-b_r) + b_r \quad (2.6.4)$$

where s is the sum of squared errors function. Although positive-part Stein estimators typically dominate the usual Stein estimators it has not been determined whether (2.6.4) dominates (2.6.2). Consider the following evidence which suggests that (2.6.4) may dominate (2.6.2). Berger and Bock (1975) consider the maximum likelihood estimator b for the orthonormal linear statistical model with three or more i.e., $b \sim N(\beta, I_K)$, $K \geq 3$. They define the spherically symmetric estimator of β to be

$$\delta(b) = h(b'b)b$$

where h is a real valued function. Under squared error loss the risk function is

$$\rho(\beta, \delta) = E\{[\delta(b) - \beta]'[\delta(b) - \beta]\}.$$

The James-Stein estimator (1961), which shrinks β toward the origin, is imbedded in the class of spherically symmetric estimators. Now, define the generalized positive rule estimator to be

$$\delta^+(b) = \{1 - g(b'b)I_{(-\infty, 0)}[h(b'b)]\}\delta(b)$$

where g is any real valued measurable function such that $1 \leq g(b'b) \leq 2$, for all $b'b$. Berger and Bock prove the risk of $\delta^+(b)$ to be less than or equal to the risk of $\delta(b)$ for all β . Thus, for a broad class of Stein estimators the positive rule which sets $h(b'b)=0$ if $h(b'b)<0$, dominates the usual Stein estimator $\delta(b)$.

Judge and Bock [(1978), p. 238] prove a similar result under weighted quadratic loss for Stein estimators of the form

$$\delta(b, s) = [I_K - h(b'Bb/s)C]b$$

where C and B are chosen such that $Q^{\frac{1}{2}}CQ^{-\frac{1}{2}}$ and $Q^{-\frac{1}{2}}BQ^{-\frac{1}{2}}$ are positive definite matrices which commute with each other and with the matrix $Q^{\frac{1}{2}}S^{-1}Q^{\frac{1}{2}}$; and, $s = (y - Xb)'(y - Xb)$.

Judge and Bock also demonstrate the risk superiority of the positive rule estimator

$$\delta^+ = [1 - as / [\sigma^2 (b - \beta_g)' S (b - \beta_g)]] \times \\ I_{[a, \infty)} [(b - \beta_g)' S (b - \beta_g) / s] (b - \beta_g) + \beta_g$$

over the regular Stein-rule estimator

$$\delta = [1 - as / [\sigma^2 (b - \beta_g)' S (b - \beta_g)]] + \beta_g$$

where β_g is a known $K \times 1$ vector. This result holds if $0 < a \leq 2 \text{tr}(S^{-1} \epsilon_L^{-2}) / (T - K + 2)$ where ϵ_L is the largest characteristic root of $S^{-1} = (X'X)^{-1}$. The estimator (2.6.4) merely replaces the constant vector β_g with the restricted least squares estimator b_r . Although it remains to be shown, given the nature of the above results it seems likely that the estimator (2.6.4) dominates (2.6.2).

(b) Mean and Covariance

The mean of the Stein-rule estimator (2.6.2) is

$$E[\delta] = \beta - a E(l_1) S^{-1} R' (R S^{-1} R')^{-1} \omega \quad (2.6.5)$$

or,

$$= \beta - a (T - K) E(1/\chi^2_{(J+2, \lambda)}) S^{-1} R' (R S^{-1} R')^{-1} \omega \quad (2.6.6)$$

where $l_1 = \chi^2_{(T-K)} / \chi^2_{(J+2, \lambda)}$. Like the pretest estimator and the restricted least squares estimator, the Stein-rule (2.6.2) is unbiased if the hypotheses are true.

The covariance matrix of the Stein-rule estimator is

$$\begin{aligned}
E\delta - E\delta' &= \sigma^2 S^{-1} \\
&- \sigma^2 S^{-1} R' (RS^{-1} R')^{-1} RS^{-1} \{2aE(l_1) - a^2 E[(l_1)^2]\} \\
&+ S^{-1} R' (RS^{-1} R')^{-1} \omega \omega' (RS^{-1} R')^{-1} RS^{-1} \times \\
&\quad \{1 - 2a[E(l_2)] + a^2 E[(l_2)^2]\} \\
&- [E(1 - a(l_1))]^2 S^{-1} R' (RS^{-1} R')^{-1} \omega \omega' (RS^{-1} R')^{-1} RS^{-1}
\end{aligned}
\tag{2.6.7}$$

where $l_2 = \chi^2_{(T-K)} / \chi^2_{(J+4, \lambda)}$.

2.6.2 Risk Under Weighted Quadratic Loss

The risk function of the estimator (2.6.2) under weighted quadratic loss is

$$\begin{aligned}
\rho(\beta, \delta; W) &= E[\delta - E\delta]' W (\delta - E\delta) = \sigma^2 \text{tr} S^{-1} W \\
&- \text{tr} \{ S^{-1} R' (RS^{-1} R')^{-1} RS^{-1} W \} [2aE(l_1) - a^2 E(l_1)^2] \\
&+ \text{tr} \{ S^{-1} R' (RS^{-1} R')^{-1} \omega \omega' (RS^{-1} R')^{-1} RS^{-1} W \} \times \\
&\quad \{ 2a[E(l_1) - E(l_2)] + a^2 [E(l_2)^2 - E(l_1)^2] \}.
\end{aligned}
\tag{2.6.8}$$

The positive constant "a" must be chosen within a specific interval for the Stein-rule to dominate the OLS estimator of β . Mittelhammer (1984) derives the upper bound, a_{\max} , for the positive shrinkage constant in the general Stein estimator (2.6.2) which ensures that $\rho(\beta, b; W) - \rho(\beta, \delta; W) = \Delta$ (where Δ is a positive semi-definite matrix) for all constraint vectors $\omega = R\beta - r$. Mittelhammer shows that a must be chosen according to

$$0 \leq a \leq [2/(T-K+2)] \times \text{tr} \{ [(RS^{-1} R')^{-1} RS^{-1} WS^{-1} R' / \lambda_L] - 2 \}
\tag{2.6.9}$$

where $S = X'X$, λ_L is the maximum characteristic root of $(RS^{-1} R')^{-1} RS^{-1} WS^{-1} R'$, and W is a positive definite weight

matrix.

2.6.3 Risk Matrix

The risk matrix of the Stein-Rule estimator (2.6.2) is denoted

$$\begin{aligned} \text{MSE}(\delta) = & E[\delta - E\delta]W(\delta - E\delta)'] = \sigma^2 S^{-1} \\ & - S^{-1}R'(RS^{-1}R')^{-1}RS^{-1}[2aE(l_1) - a^2E(l_1)^2] \\ & + S^{-1}R'(RS^{-1}R')^{-1}\delta\delta'(RS^{-1}R')^{-1}RS^{-1} \times \\ & \{2a[E(l_1) - E(l_2)] + a^2[E(l_2)^2 - E(l_1)^2]\}. \end{aligned} \quad (2.6.10)$$

Notice that $\text{MSE}(b) - \text{MSE}(\delta) = \Delta$ (a positive semi-definite matrix) when (2.6.9) is satisfied.

2.7 Stein-Rule Problems and Alternatives

The classical least squares estimator of β in (2.1.1) is $b = S^{-1}X'y$ and is minimax with constant risk $\rho(\beta, b; W) = \text{tr} S^{-1}W$ under weighted quadratic loss. The Stein-rule estimator δ dominates b for all β when $K \geq 3$, provided that the other design related conditions are met. The Stein estimator is able to improve upon the OLS estimator in terms of risk because of the way it uses prior information. The basic fact remains, however, that $\rho(\beta, \delta; W)$ is substantially less than $\rho(\beta, b; W)$ over a relatively small portion of the parameter space.

By using (2.6.2) to estimate β one is implicitly speculating that linear combinations of the true parameters lie near or within the ellipsoid

$$C = \{\beta : (R\beta - r)'(RS^{-1}R')^{-1}(R\beta - r) \leq p\}$$

where the distance between $R\beta$ and r is measured in the

$(RS^{-1}R')^{-1}$ metric and p is a known constant. The matrix $(RS^{-1}R')^{-1}$ essentially determines the shape of the ellipsoid. The more tightly one can 'draw' the ellipsoid and the more accurately one can define the restrictions, the better the potential risk performance of the resulting estimator. However, if the ellipsoid is drawn in a region which is not 'near' the true point, then the potential risk improvement may be very small. So, the first problem is that of deciding upon the prior information to incorporate into (2.6.2).

Second, even if the prior information is good, it is still quite possible for an individual element(s) of δ to be estimated rather poorly using δ ; in other words, it is possible for individual elements of δ to have higher risk than corresponding elements in b . Efron and Morris (1972) call this **component risk** and show that in many instances it can be quite large.

Several researchers have addressed this problem. Efron and Morris (1972) suggest a "limited translation empirical Bayes estimator" which combines features of maximum likelihood, James-Stein, and Bayes estimators. Using this estimator, Efron and Morris (1972) show that under squared error loss substantial gains in component risk are possible without much sacrifice in ensemble risk. Stein (1981) has proposed another, more robust, estimator based on order statistics which achieves similar results and is also minimax under squared error loss.

A third major problem with using a Stein estimator like (2.6.2) is the difficulty in performing hypotheses tests. The sampling distribution of the statistic δ is uncertain. In addition, its covariance matrix (2.6.7) contains unknown parameters; if these are replaced with estimates, the estimated covariance matrix will have an unknown sampling distribution. Given these circumstances, the usual hypothesis tests, which are based on the likelihood ratio or Wald principle, cannot be performed.

One possible solution to this problem is to abandon the classical framework altogether and adopt a Bayesian or empirical Bayesian approach. Berger (1980) develops a generalized robust Bayes estimator which is in form similar to the James-Stein (1961) estimator. Berger's estimator permits great flexibility in incorporating prior information, it is robust with respect to misspecification, its prior density is expressible in closed form, and it is admissible over important regions of the parameter space. Most importantly, Berger's estimator permits calculation of the covariance matrix of the posterior distribution, given δ and the prior density.

Although the statistical properties of these alternatives to Stein-rule estimation will not be featured in subsequent chapters, the robust generalized Bayes and new-Stein estimators are important because each addresses a perceived weakness of the regular Stein-rule. In Chapter 7 these techniques will resurface as results are summarized

and future research planned.

2.8 Confidence Sets and Hypothesis Tests

In this section, the construction of confidence ellipsoids under normal distribution theory will be summarized (following Scheffe 1959) and related to the theory of hypothesis testing.

Consider again, the linear model (2.1.1)

$$y = X\beta + e \quad (2.1.1)$$

assuming further that $e \sim N(0, \sigma^2 I_T)$. The model may be concisely summarized as

$$y \sim N(X\beta, \sigma^2); \quad \text{rank}(X) = K. \quad (2.8.1)$$

Now suppose that the distribution of the observed random variables y is completely determined by the values of the unknown parameters $\theta = \{\theta_1, \theta_2, \dots, \theta_m\}$ and that $\psi = \{\psi_1, \psi_2, \dots, \psi_q\}$ are specified functions of the parameters. Geometrically, ψ is a point in q -dimensional space with coordinates $\{\psi_1, \psi_2, \dots, \psi_q\}$, θ is a point in m -dimensional space with coordinates $\{\theta_1, \theta_2, \dots, \theta_m\}$, and y is a point in T -dimensional space with coordinates $\{y_1, y_2, \dots, y_T\}$. Now suppose that for every possible point in the sample space y , a region $R(y)$ in the q -dimensional ψ -space is determined. If the probability that $R(y)$ covers the true point ψ is a pre-assigned constant $1-\alpha$, then $R(y)$ is called a confidence set for ψ with confidence coefficient $1-\alpha$.

This relationship is denoted

$$\Pr[y: \psi(\theta) \in R(y) | \theta] = 1-\alpha \quad \text{for all } \theta \in \Theta. \quad (2.8.2)$$

Notice that the probability holds for any value of θ in the

parameter space Θ . The classical statistician is interested in the long-run proportion $1-\alpha$ of the calculated confidence sets covering the true value of $\psi(\theta)$ being estimated.

Bayesians, who consider the underlying parameters θ to be random (with respect to the prior distribution), would interpret $\psi(\theta)$ as random and concern themselves with estimation of the posterior p.d.f. The probability that $\psi(\theta)$ lies in a subregion R of the ψ -space would be

$$\Pr[\psi(\theta) \in R | y] = \int_R p(\psi(\theta) | y) d\theta \quad (2.8.3)$$

where $p(\psi(\theta) | y)$ is the posterior p.d.f. of $\psi(\theta)$ given the sample y . The probability in (2.8.3) measures the statistician's degree of belief that $\psi(\theta)$ lies in the region R given the sample and prior information. Fixing the probability at $1-\alpha$ such that (2.8.3) holds, yields the Bayesian equivalent to (2.8.2).

Returning to the classical interpretation it is possible to show that the usual F-test based on the likelihood ratio principle is an application of this rather general theory of confidence sets. From section 2.4, recall the normal linear statistical model subject to J independent linear hypotheses described in equation (2.4.1) and, using (2.8.1), suppose one wishes to test the null hypothesis $R\beta - r = 0$ against the alternative $R\beta - r \neq 0$, where R is $J \times K$ of rank J and r is a $J \times 1$ vector of known constants.

Let $\{\theta_1, \theta_2, \dots, \theta_m\} = \{\beta_1, \beta_2, \dots, \beta_K, \sigma^2\}$ and

$$\hat{\psi} \sim N(\psi, A) = R\beta \sim N(R\beta, \sigma^2 R S^{-1} R').$$

Using results from normal distribution theory [Schmidt (1976), Lemma 2, p. 11 and Theorem 10, p. 22] one can form the ellipsoid

$$\begin{aligned} & (\hat{\psi} - \psi)' A^{-1} (\hat{\psi} - \psi) \sim \chi^2_q \\ \text{or,} \quad & (Rb - R\beta)' (RS^{-1}R')^{-1} (Rb - R\beta) \sim \sigma^2 \chi^2_J. \end{aligned} \quad (2.8.4)$$

Replacing σ^2 by its unbiased estimator $\hat{\sigma}^2$ and recalling that $\hat{\sigma}^2$ and b are statistically independent, (2.8.4) can be rewritten as

$$(Rb - R\beta)' (RS^{-1}R')^{-1} (Rb - R\beta) / \hat{\sigma}^2 \sim F_{J, T-K}. \quad (2.8.5)$$

Under the null hypothesis, $R\beta = r$ and (2.8.5) becomes

$$(Rb - r)' (RS^{-1}R')^{-1} (Rb - r) / \hat{\sigma}^2 \sim F_{J, T-K, \lambda}. \quad (2.8.6)$$

where $\lambda = (R\beta - r)' (RS^{-1}R')^{-1} (R\beta - r) / 2\sigma^2$. Expression (2.8.6) is the familiar test statistic associated with the null hypothesis $R\beta - r = 0$.

A $100(1-\alpha)\%$ confidence set (or confidence ellipsoid) can be obtained using the fact that the desired probability of the F-random variable in (2.8.5) being less than or equal to $F_{J, T-K}$ is $1-\alpha$. This yields

$$(Rb - R\beta)' (RS^{-1}R')^{-1} (Rb - R\beta) \leq \hat{\sigma}^2 F_{\alpha; J, T-K} \quad (2.8.7)$$

or,

$$(b - \beta)' R' (RS^{-1}R')^{-1} R (b - \beta) \leq \hat{\sigma}^2 F_{\alpha; J, T-K}. \quad (2.8.8)$$

The inequality (2.8.7) determines the confidence ellipsoid in the J -dimensional ψ -space centered at a linear transformation of the OLS estimates Rb , whereas (2.8.8) determines a confidence ellipsoid in the K -dimensional parameter space centered at the OLS estimates themselves. Although the two ellipsoids are measured in spaces of

different dimension and in different metrics, they are equivalent (both metric spaces are rank J). The probability that the ellipsoid (2.8.8) covers the true parameter point is $1-\alpha$ regardless of the actual values of β and σ^2 .

Chapter 3 Statistical Models and Methods

3.1 Time Series Models

3.1.1 Stochastic Processes

- (a) Autoregressive Processes
- (b) Moving Average Processes
- (c) Autoregressive Integrated Moving Average Processes

3.1.2 Univariate Models

- (a) Identification
- (b) Estimation
- (c) Diagnostic Checking

3.1.3 Distributed Lag Models

- (a) Finite Distributed Lags of Unknown Length
- (b) Estimating Lag Length
- (c) Model Selection Criteria
- (d) Estimating Lag Length for More Than One Variable

3.1.4 Polynomial Distributed Lag Models

- (a) Polynomial Restrictions
- (b) Model Selection
- (c) Effects of Misspecification
- (d) Forecast Evaluation

3.2 Nonlinear Models

3.2.1 Statistical Model

- (a) Statistical Properties
- (b) Tests of Hypotheses

3.2.2 Numerical Techniques

- (a) Newton-Raphson
- (b) Method of Scoring
- (c) Guass-Newton

3.2.3 Example: The Probit Regression Model

3.3 Generalized Linear Models

3.3.1 Estimating GLIM

- (a) Components of GLIM
- (b) Likelihood Function for the GLIM
- (c) Algorithm for Estimating GLIM

3.3.2 Example: The Probit Regression Model Revisited

3.4 Computer Intensive Research Techniques

3.4.1 Monte Carlo

3.4.2 The Bootstrap

Chapter 3 Statistical Models and Methods

This chapter contains discussions on times series models, nonlinear models and their estimation, generalized linear models and computer intensive research techniques.

3.1 Time Series Models

A time series consists of a set of observations on a random variable y taken at equally spaced intervals over time. Each random variable y_t has a mean μ_t and a zero mean random component e_t . Model (2.1.1) can be thought of in these terms if y_t is taken to be a time series with mean $\mu_t = x_t' \beta$ where x_t is the t^{th} row of the matrix of explanatory variables X . Sometimes, however, the researcher is either unable or unwilling to specify an explanatory model for μ . In such cases univariate techniques have been proposed which specify autoregressive moving average models for stationary time series. These techniques, unified by Box and Jenkins (1976), have been used with some success for short-run forecasting of economic time series. For longer time series (those of at least 40 to 50 observations), Granger and Newbold (1977) find the Box-Jenkins approach to be particularly valuable when the series has proved difficult to predict by routine methods.

In this portion of the chapter, the following issues will be discussed: stochastic processes, univariate models, and the polynomial distributed lag explanatory model (i.e., where $\mu = X\beta$ and the finite lag weights β_i are

assumed to fall along an r^{th} degree polynomial).

3.1.1 Stochastic Processes

A stochastic process has been described as "a statistical phenomenon that evolves in time according to probabilistic laws." [Granger and Newbold (1977), p. 33] Mathematically, a stochastic process is defined to be a collection of random variables $\{y_t, t \in T\}$, where T denotes the set of time points at which the process is observed. The random variable is usually discrete, implying $t=0, \pm 1, \pm 2, \dots$.

An important point to remember is that each of the observed variables in the series represents a different random variable, each having its own p.d.f. Thus, given a sample, one has only a single observation of each random variable at each time t . Therefore, the observed time series is actually a single realization of the stochastic process. Given this fact, an obvious way to describe a stochastic process is by its joint probability density function $f(y_1, y_2, \dots, y_T)$. This method is difficult to use; instead, researchers choose to describe the process by the first few moments of the probability distribution. But, even an investigation of the first two moments of the joint p.d.f., (the means, variances, and covariances) is impossible without making further assumptions.

An important assumption which allows the researcher to make inferences using the mean, variances, and covariances of a stochastic process is that of

stationarity. A stochastic process is defined to be weakly stationary if and only if

- i. $E(y_t) = \mu < \infty$ for all t
- ii. $E[(y_t - \mu)^2] < \infty$ for all t
- iii. $E[(y_t - \mu)(y_{t+k} - \mu)] = \gamma(k)$ for all t, k .

Thus, the mean and variances must be finite and the autocovariance function $\gamma(k)$ must not depend on the time period t (only on the distance between two points over time). Another type of stationarity is often cited in the time series literature. Strict stationarity requires that the multivariate distribution of $(y_1, y_2, \dots, y_{t+k})$ be identical to that of the time shifted set $(y_s, y_{s+1}, \dots, y_{s+k})$ for all t, s , and k . Verification of strict stationarity is difficult because precise knowledge of the joint p.d.f. seldom exists. Note, however, if the stochastic process is normally distributed, then two moments completely describe its probability density; therefore using the definitions developed above, weak stationarity and strong stationarity will be equivalent.

Recall that in models discussed in Chapter 2, it was assumed that the sample observations are statistically independent and identically distributed. The joint p.d.f. can be found by taking the product of T density functions. In time series analysis, independence of sample observations is not assumed. Actually, it is the functional relationship among the variables of a time series which allows a model to be built. The linear

relationship between two random variables in the time series is measured by the covariance or autocovariance function. The theoretical autocovariance function of the random sequence y_t is defined to be

$$\gamma(k) = E[(y_t - Ey_t)(y_{t+k} - Ey_{t+k})] \quad (3.1.1)$$

If the process is weakly stationary, then $Ey_t = Ey_{t-k} = \mu$ and

$$\gamma(k) = E[(y_t - \mu)(y_{t+k} - \mu)] \quad k=0,1,2,\dots \quad (3.1.2)$$

This quantity is usually normalized by dividing through by $\gamma(0) = \text{Var}(y_t)$. This result is called the theoretical autocorrelation function $\rho(k)$ of the random sequence y_t and is denoted

$$\rho(k) = \gamma(k)/\gamma(0) \quad k=1,2,\dots \quad (3.1.3)$$

Another function used to describe a stochastic process is the partial autocorrelation function. The k^{th} partial autocorrelation coefficient measures the correlation between y_t and y_{t-k} given $y_{t-1}, y_{t-2}, \dots, y_{t-k+1}$. The actual derivation of the partial autocorrelation functions is presented in the next section.

In summary, if the stochastic process is nonstationary, then it must be made stationary before attempting to fit a model. Once this is done, it can be identified and described by certain linear relationships between any two of the observations.

(a) Autoregressive Processes

Without loss of generality it is assumed that y_t has zero mean, or equivalently, if y_t^* is some other weakly stationary time series, then $y_t = y_t^* - Ey_t^*$. Given this, a

finite autoregressive process of order p is denoted $AR(p)$ and has the form

$$y_t = \theta_1 y_{t-1} + \theta_2 y_{t-2} + \dots + \theta_p y_{t-p} + e_t \quad (3.1.4)$$

or,

$$(1 - \theta_1 L - \theta_2 L^2 - \dots - \theta_p L^p) y_t = e_t \quad (3.1.5)$$

where L is the lag operator $L^i y_t = y_{t-i}$, $E(e_t) = 0$, and $\text{Var}(e_t) = \sigma^2$ for all t . Equation (3.1.5) can also be expressed as

$$\theta(L) y_t = e_t \quad (3.1.6)$$

where $\theta(L) = (1 - \theta_1 L - \theta_2 L^2 - \dots - \theta_p L^p)$ is a polynomial of degree p in the lag operator. The autoregressive process (3.1.6) is stationary if the solutions to the difference equation

$$1 - \theta_1 z - \theta_2 z^2 - \dots - \theta_p z^p = 0 \quad (3.1.7)$$

lie outside the complex unit circle. Stationarity implies that lagged effects become smaller the further in the past they occurred.

(b) Moving Average Processes

A moving average process of order q is denoted $MA(q)$ and has the form

$$y_t = e_t + \alpha_1 e_{t-1} + \alpha_2 e_{t-2} + \dots + \alpha_q e_{t-q} \quad (3.1.8)$$

or,

$$y_t = (1 + \alpha_1 L + \dots + \alpha_q L^q) e_t \quad (3.1.9)$$

$$= \alpha(L) e_t \quad (3.1.10)$$

where e_t has zero mean and finite variance σ^2 . The MA process y_t is stationary if $\sigma^2 \alpha(L)$ is finite.

Not all stationary MA operators can be inverted. To

guarantee invertibility of an MA(q) process, the solutions to

$$1 + \alpha_1 z + \alpha_2 z^2 + \dots + \alpha_q z^q = 0 \quad (3.1.11)$$

must lie outside the complex unit circle. Invertibility implies that past values of y have a decreasing effect on current values of y .

(c) Autoregressive Integrated Moving Average Processes

It is possible to generalize the two schemes discussed above by combining them into an autoregressive moving average process of order (p, q) , often denoted ARMA(p, q). This process can be expressed as

$$y_t = \theta_1 y_{t-1} + \dots + \theta_p y_{t-p} + e_t + \alpha_1 e_{t-1} + \dots + \alpha_q e_{t-q} \quad (3.1.12)$$

or,

$$\theta(L)y_t = \alpha(L)e_t. \quad (3.1.13)$$

Stationarity and invertibility of (3.1.13) require that

$$\alpha(L) = 1 + \alpha_1 z + \alpha_2 z^2 + \dots + \alpha_q z^q \neq 0 \quad |z| \leq 1 \quad (3.1.14)$$

$$\theta(L) = 1 - \theta_1 z - \theta_2 z^2 - \dots - \theta_p z^p \neq 0 \quad |z| \leq 1 \quad (3.1.15)$$

If the centered series y_t (centered about its mean) is not stationary, Box and Jenkins suggest (1) differencing, (2) applying a suitable transformation (e.g., $\ln(y_t)$) or (3) transforming and differencing the time series until stationarity is achieved. Once stationarity is induced in this manner, the resulting process is called an autoregressive integrated moving average process or an ARIMA(p, d, q) where d is the number of first differences taken to achieve weak stationarity. The resulting process

is denoted

$$\theta(L)(1-L)^d y_t = \alpha(L)e_t \quad (3.1.16)$$

where $\theta(L)$ is of degree p and $\alpha(L)$ is of degree q .

There is an important dual relationship between the autoregressive process and the moving average process. Wold's decomposition theorem (1954) can be used to show that a stationary AR process can be represented by an MA process of infinite degree. The resulting process is referred to as the moving average representation of the autoregressive process of order p (see equation 3.1.35). Likewise, any MA(q) process can be represented by an infinite AR process. The importance of this duality will become apparent in the following discussion of univariate modeling.

3.1.2 Univariate Modeling

In the previous section the ARIMA(p,d,q) stochastic process was defined and the stationarity and invertibility conditions necessary for its identification were presented. It was assumed that the time series under consideration could be represented by a model from this class after removal of any deterministic component, including a non-zero mean, and/or the application of some suitable transformation of the data.

The basic strategy for construction ARIMA models is based on a 3 step iterative procedure involving:

- (1) model identification
- (2) model estimation

(3) diagnostic checking.

In the following 3 subsections, each of these steps will be discussed in some detail.

(a) Model identification

In order to identify a particular model from the class (3.1.16) one must choose p, q , and d . This part of the model building process is anything but precise and requires both patience and experience on the part of the investigator.

The first step in identifying the model is to check the stationarity of the time series. At this stage, there is no good substitute for first plotting the data points. The degree of differencing required to ensure stationarity can usually be determined by inspecting the plots. A linear trend can be removed by first differencing and a quadratic trend by second differencing. If the variance of the series appears to increase proportionately with the mean, then a logarithmic transformation may be required. In sophisticated applications, the Box-Cox (1964) transformation may be used. The logarithmic transformation is a special case of the Box-Cox transformation. For a given λ , the transformed value, y_t^λ , is given by

$$y_t^\lambda = \begin{cases} (y_t^\lambda - 1) / \lambda & \lambda \neq 0 \\ \ln(y_t) & \lambda = 0 \end{cases} \quad (3.1.17)$$

For details and examples of this approach to inducing stationarity, consult Nelson and Granger (1979).

Once the series has been appropriately transformed

into a stationary time series, one examines the sample autocorrelation and partial autocorrelation functions for clues as to the choice of q and p for the moving average and autoregressive operators.

The two most important tools at the disposal of the investigator during this stage of the model building process are the sample autocorrelation and partial autocorrelation functions. For a stationary time series y_t , the k^{th} theoretical autocovariance function of an AR(p) process can be derived from the moving average representation and is denoted

$$\gamma(k) = \sigma^2 \sum_{i=0}^{\infty} \alpha_i \alpha_{i+k} \quad \text{where } \alpha_0 = 1 \quad (3.1.18)$$

provided $\sum |\alpha_i|$ converges. The corresponding autocorrelation function $\rho(k)$ can be obtained by taking $\gamma(k)/\gamma(0)$. Similarly, the k^{th} theoretical autocorrelation function for an invertible MA(q) process is given by

$$\rho(k) = \begin{cases} 1 & k=0 \\ \frac{\sum_{i=0}^{q-k} \alpha_i \alpha_{i+k}}{\sum_{i=0}^q \alpha_i^2} & k=1, \dots, q \\ 0 & k > q \\ \rho(-k) & k < 0. \end{cases} \quad (3.1.19)$$

Notice that the autocorrelation function of the invertible MA(q) process is zero for all $k > q$ and that it declines toward zero only as k approaches infinity for the stationary AR(p) process.

As previously stated, estimates of both

autocorrelation and partial autocorrelation functions are needed in order to identify a time series. Although several procedures have been suggested [see Jenkins and Watts (1968)], Box and Jenkins find the most satisfactory estimate of the k^{th} autocorrelation $\rho(k)$ to be

$$r(k) = c(k)/c(0)$$

where

$$c(k) = T^{-1} \sum_{t=1}^{T-K} (y_t - \bar{y})(y_{t+k} - \bar{y}) \quad k=0,1,2,\dots,K$$

is the estimate of the autocovariance $\gamma(k)$, and \bar{y} is the mean of the stationary time series. Note, this estimator of $\gamma(k)$ is biased and therefore so is the estimator $r(k)$. In general, $c(k)$ and $r(k)$ have lower risk under squared error loss than the unbiased estimators of $\rho(k)$ and $\gamma(k)$. In order to derive the partial autocorrelation functions, pre-multiply the following AR(p) process by y_{t-k}

$$y_t = \theta_1 y_{t-1} + \dots + \theta_p y_{t-p} + e_t$$

to obtain

$$y_{t-k} y_t = \theta_1 y_{t-k} y_{t-1} + \dots + \theta_p y_{t-k} y_{t-p} + y_{t-k} e_t. \quad (3.1.20)$$

The expectation $E(y_{t-k} e_t) = 0$ for $k > 0$, since y_{t-k} can only be influenced by random shocks which precede it. Thus, taking the expectation of (3.1.20) yields the difference equation

$$\gamma(k) = \theta_1 \gamma(k-1) + \theta_2 \gamma(k-2) + \dots + \theta_p \gamma(k-p) \quad k > 0 \quad (3.1.21)$$

which, if divided through by $\gamma(0)$, yields

$$\rho(k) = \theta_1 \rho(k-1) + \theta_2 \rho(k-2) + \dots + \theta_p \rho(k-p) \quad k > 0. \quad (3.1.22)$$

For an autoregressive process of order p , one obtains a set

of linear equations of the form

$$\begin{aligned}\rho(1) &= \theta_1 \rho(0) + \theta_1 \rho(1) + \dots + \theta_p \rho(p-1) \\ \rho(2) &= \theta_1 \rho(1) + \theta_1 \rho(0) + \dots + \theta_p \rho(p-2) \\ &\vdots \\ \rho(p) &= \theta_1 \rho(p-1) + \theta_1 \rho(p-2) + \dots + \theta_p \rho(0).\end{aligned}$$

This set of linear equations is called the Yule-Walker Equations and may be written in matrix form

$$\rho = \rho_p \theta \quad (3.1.23)$$

where $\rho = (\rho(1), \dots, \rho(p))'$, $\theta = (\theta_1, \dots, \theta_p)'$, $\rho(0)=1$, and

$$\rho_p = \begin{bmatrix} 1 & \rho(1) & \dots & \rho(p-1) \\ \rho(1) & 1 & \rho(1) & \rho(p-2) \\ \vdots & & \ddots & \vdots \\ \rho(p-1) & \rho(p-2) & \dots & 1 \end{bmatrix}.$$

The i^{th} equation of (3.1.23) is

$$\rho(i) = \theta_1 \rho(i-1) + \dots + \theta_p \rho(p-i)$$

where θ_i , $i=1,2,\dots,p-1$ is the i^{th} coefficient of an AR(p) process and θ_p is the last coefficient. If ρ_p^{-1} exists, then (3.1.23) can be solved for θ .

The quantity θ_p is regarded as a function of the lag p and is called the partial autocorrelation function. For an autoregressive process of order p , the k^{th} partial autocorrelation function θ_k will be nonzero for $k \leq p$ and zero for $k > p$. The partial autocorrelation function can be estimated by successively fitting autoregressive processes of orders $1,2,3,\dots$ by ordinary least squares and picking out the coefficients of the last term in the estimated

equation at each stage.

Whereas the autocorrelation function of an $AR(p)$ process tails off as $k \rightarrow \infty$, its partial autocorrelation function becomes zero after the p^{th} lag. Conversely, the autocorrelation function of the $MA(q)$ process cuts off after lag q , while its partial autocorrelation function declines toward zero in the limit. If both autocorrelations and partial autocorrelations tail off, a mixed process containing a p^{th} order autoregressive component and a q^{th} order moving average component is suggested. The autocorrelation function in this case exhibits a mixture of exponential and damped sine waves after the first $q-p$ lags and the partial autocorrelation function for the mixed process is dominated by a mixture of exponentials and damped sine waves after the first $p-q$ lags. Box and Jenkins (1976) summarize these findings on page 79.

Identification of the series is not as easy as this might suggest. Estimated autocorrelations tend to have large variances and to be highly correlated with one another. As Kendall (1945) notes, a strict correspondence between estimated and theoretical autocorrelation functions cannot be expected. In general, it is only possible to gain knowledge of the broad characteristics of the series by examining the sample autocorrelation and partial autocorrelation functions. Usually, several models must be entertained and carried through to the estimation and

diagnostic checking stages.

Given the difficulty of approximating the theoretical autocorrelation and partial autocorrelation functions with sample counterparts, it is important to gain some idea of how far apart the two might actually be. Box and Jenkins use variance estimates as "informal guides" to aid in the determination of how far the estimated value is from its theoretical value.

The variance of the estimated autocorrelation $r(k)$ at lags $k > q$ can be approximated by

$$\text{Var}(r(k)) \cong T^{-1} \{1 + 2[r(1)^2 + r(2)^2 + \dots + r(q)^2]\} \quad (2.1.24)$$

Anderson (1942) shows that for moderate T , the distribution of the estimated autocorrelation function coefficient, whose theoretical value is zero, is approximately normal. Thus, under the null hypothesis that $\rho(k)=0$ (against the alternative $\rho \neq 0$) the estimate $r(k)$ divided by $[\text{Var}(r(k))]^{\frac{1}{2}}$ will be distributed approximately $N(0,1)$.

For the partial autocorrelation function, Quenouille (1949) showed that under the null hypothesis that the process is $AR(p)$, the estimated autocorrelations of order $p+i$, $i > 0$, are approximately independently distributed and that

$$\hat{\sigma}[\hat{\theta}_k] \cong T^{-\frac{1}{2}} \quad k \geq p+1 \quad (3.1.25)$$

[see Box and Jenkins (1976), pp. 34-35, 65, and 177-178].

(b) Estimation

Having tentatively identified d , p , and q , efficient

estimates of the parameters must be obtained. One possible approach is to minimize the sum of squared errors of the model (3.1.16)

$$\theta(L)(1-L)^d y_t = \alpha(L)e_t$$

or,

$$\theta(L)z_t = \alpha(L)e_t \quad (3.1.26)$$

where z_t is the stationary, time series $(1-L)^d y_t$. Least squares estimation of this model presents two immediate difficulties:

- (i) the equations will contain unknown starting values $z^*=(z_0, z_1, \dots, z_{1-p})$ and $e^*=(e_1, e_2, \dots, e_{1-q})$ and
- (ii) the model is nonlinear in the parameters.

Below, we consider both conditional and unconditional methods of estimating a univariate ARIMA model.

Conditional Approach

One way of overcoming the starting value problem is to replace z^* and e^* with reasonable assumed values; estimation is then conditional on the assumed values z^* and e^* . A possible choice of starting values is $E(z^*)$ and $E(e^*)$. Given the stationary series z_t , a normal log likelihood function with parameters $(\theta, \alpha, \sigma^2)$, and the assumed starting values (z^*, e^*) the conditional log likelihood function is

$$\ell_* = \ln L_*(\theta, \alpha, \sigma^2) \propto -n \ln(\sigma) - s_*(\theta, \alpha) / 2\sigma^2 \quad (3.1.27)$$

where $s_*(\theta, \alpha) = \sum_{t=1}^T e_t^2(\theta, \alpha | z^*, e^*, z)$. The star subscript

reflects the fact that the likelihood and sum of squares functions are conditional on the assumed starting values

(z^*, e^*) . The maximum likelihood estimates are those values of θ, α , and σ which maximize the conditional likelihood function (3.1.27).

Similarly, one could obtain a conditional nonlinear least squares estimator of (θ, α, σ) by minimizing $s_*(\theta, \alpha, \sigma)$. In either case, some sort of numerical optimization technique must be employed. These techniques will be discussed in Section 3.2.3. Whatever algorithm is used, successive calculation of the residuals $e_t(\theta, \alpha | z^*, e^*, z)$, $t=1, 2, \dots, T$, is required for use in the sum of squares function; furthermore, as each new round of estimates for θ and α is calculated, another set of e_t must be generated and ℓ_* or s_* minimized again. [Box and Jenkins, pp. 209-211, (1976)].

Unconditional Approach

It is also possible to derive unconditional estimates of the parameters θ and α . Box and Jenkins [(1976), p. 213] show that the unconditional log likelihood function is given by

$$\ell(\theta, \alpha, \sigma^2) = f(\theta, \alpha) - T \ln(\sigma) - s(\theta, \alpha)/2\sigma^2 \quad (3.1.28)$$

where $f(\theta, \alpha)$ is a function of θ and α . The unconditional sum of squares is $s(\theta, \alpha) = \sum (e_t | \theta, \alpha, z)^2$ where $(e_t | \theta, \alpha, z) = E[(e_t | \theta, \alpha, z)]$. For large T , $f(\theta, \alpha)$ is dominated by $s(\theta, \alpha)/2\sigma^2$ and the estimates obtained by minimizing $s(\theta, \alpha)$ will be approximately equivalent to the maximum likelihood estimates. Other procedures, such as Marquardt's (1963) compromise, have been used to find efficient estimates of

the parameters θ, α , and σ^2 .

For a stationary, invertible Gaussian process the maximum likelihood estimator is consistent, asymptotically efficient, and asymptotically normally distributed. The inverse of Fisher's information matrix evaluated at the MLE's can be used as an estimator of the covariance matrix, $\text{Cov}(\theta, \alpha, \sigma^2)$. For a discussion of maximum likelihood estimation of an ARMA(p,q) process consult Newbold (1974), Anderson (1977), or Ansley (1979).

(c) Diagnostic Checking

Once the model has been identified and estimated, the adequacy of the fitted model should be checked. If one finds evidence that the model inadequately represents the time series, then one adaptively identifies an alternative model and re-estimates its parameters. Once satisfied with the model, forecasts can be generated.

A successful univariate time series model should capture systematic movements in the data with a minimum number of parameters. If this is achieved, then the residuals of the model will be a white noise random process. Again, as Box and Jenkins [(1976), p. 289] note, there is no substitute for visually inspecting the plot of the residuals as an initial step in the diagnostic checking stage of univariate model building. If any pattern can be detected in the plot, then the fitted model is probably poor.

Another technique used for diagnostic checking is

based on overfitting the model. Having identified what is believed to be the true model, one fits another model containing additional parameters covering suspected inadequacies. Assume for the moment that an ARMA(p,q) model is being compared to an ARMA(p+j,q) model. Under the null hypothesis $\theta_{p+1}=\theta_{p+2}=\dots=\theta_{p+j}=0$ and under the alternative hypothesis $\theta_{p+1}\neq 0, \theta_{p+2}\neq 0, \dots, \theta_{p+j}\neq 0$. The likelihood ratio test takes the form

$$u = -2\ln[L_0(\theta, \alpha, \sigma^2)/L_1(\theta, \alpha, \sigma^2)] \approx \chi_j^2 \quad \text{if } H_0: \text{ true} \quad (3.1.29)$$

where L_0 and L_1 denote the maximized value of the likelihood functions under null and alternative hypotheses, respectively.

It is also possible to compare ARMA(p,q) and ARMA(p+j,q) processes without having to estimate the larger model using a Lagrange multiplier (LM) test [Harvey (1983a)]. For the LM test the residuals e_t from the ARMA(p,q) model are regressed on the full set of p+j+q derivatives evaluated under the null hypothesis. The resulting statistic is $TR^2 \approx \chi_j^2$, if the null hypothesis is true. Note that a Lagrange multiplier or likelihood ratio test could also be performed in an obvious way by augmenting the MA component of the ARMA(p,q) process, yielding an ARMA(p,q+j), and proceeding as discussed for the ARMA(p+j,q).

Another approach to diagnostic checking makes use of the residuals from the fitted ARMA(p,q) model. These have

been used in several ways.

First, if e_1, e_2, \dots, e_T is a sequence from a white noise process, then for moderately large T , the sample autocorrelations are uncorrelated and normally distributed with variance $1/T$. Having fit the model $\theta(L)z_t = \alpha(L)e_t$ by the method of maximum likelihood and denoting the ML estimates $(\hat{\theta}, \hat{\alpha})$, the fitted model can be written

$$\hat{e}_t = \hat{\alpha}^{-1}(L)\hat{\theta}(L)z_t.$$

If the model is adequate, then $\hat{e}_t = e_t + O(T^{-1/2})$ [see Box and Jenkins (1976), p. 289]. As a result, for longer series it is reasonable to expect the residual autocorrelations

$$r_k(\hat{e}) = \frac{\sum_{t=1+k}^T \hat{e}_t \hat{e}_{t-k}}{\sum_{t=1}^T \hat{e}_t^2} \quad (3.1.30)$$

to yield valuable information about the adequacy of the fitted model. Using this fact and a result from Anderson (1942), the model is considered to be adequate if each of the $r_k(\hat{e})$ falls within the $\pm T^{-1/2}$ interval. If a predetermined number of the residual autocorrelations fall outside the interval, then the fitted model should be reconsidered. Caution must be used, however, since the asymptotic standard deviations for small k may be much less than $T^{-1/2}$. For low order lags, use of $T^{-1/2}$ as the standard error may cause one to underestimate the significance of departures from zero of the autocorrelations [Box and Jenkins (1976) p. 290].

Box and Pierce (1970) provide a useful test of the

residual autocorrelation functions. The Box-Pierce test is based on the first M autocorrelations and is computed using

$$Q = T \sum_{k=1}^M \hat{r}_k^2(\hat{e}) \quad (3.1.31)$$

which is distributed approximately as a central chi-square random variable with $M-q-p$ degrees of freedom under the null hypothesis that \hat{e} is white noise. According to Granger and Newbold [(1977), p. 93] the validity of this test relies on M being moderately large, i.e., at least 20.

Harvey prefers a similar statistic [Harvey (1983b), p. 148]

$$Q^* = T(T+2) \sum_{k=1}^M (T-k)^{-1} \hat{r}_k^2. \quad (3.1.32)$$

This statistic is referred to as the modified Box-Pierce or the Box-Ljung statistic and is tested using the χ^2_{M-p-q} distribution.

Finally, some authors use one or more of the so-called model selection criteria to gain insight into the adequacy of nonnested models. These criteria are discussed in some detail in the next portion of the chapter as attention is focused on the fitting of polynomial distributed lag models.

(d) Forecasting

Consider again the model

$$\theta(L)z_t = \alpha(L)e_t \quad (3.1.33)$$

where z_t is a weakly stationary time series of length T .

Let

$$[\alpha(L)/\theta(L)]e_t = \psi(L)e_t \quad (3.1.34)$$

and rewrite the model (3.1.33) using (3.1.34) as

$$z_t = \psi(L)e_t$$

which if expanded becomes

$$z_t = \sum_{j=0}^{\infty} \psi_j e_{t-j}. \quad (3.1.35)$$

This last expression simply means that the current value of the series z_t can be represented by a linear function of all past random disturbances e_t where $t \in (-\infty, t]$. In other words, this is the moving average representation of the stationary time series z_t referred to in section 3.1.1.(c).

Let the forecasted value of z_{t+1} be denoted

$$\hat{z}_t(1) = \sum_{j=0}^{\infty} \psi_j e_{t+1-j} \quad (3.1.36)$$

and notice that the first $l-1$ terms in the series will be zero since future error terms, e_{t+i} $i > 0$, do not exist as of time t .

Suppose that the best forecast is denoted

$$\hat{z}_t(1) = \psi^* e_t + \psi_{1+1}^* e_{t-1} + \psi_{1+2}^* e_{t-2} + \dots \quad (3.1.37)$$

where the lag weights $\psi_1^*, \psi_{1+1}^*, \dots$ are to be determined in an optimal way. The mean square error of the forecast is

$$\begin{aligned} E[z_{t+1} - \hat{z}_t(1)]^2 = \\ (1 + \psi_{1+1}^2 + \dots + \psi_{1+l-1}^2) \sigma^2 + \sum_{j=0}^{\infty} [\psi_{1+j} - \psi_{1+j}^*]^2 \sigma^2 \end{aligned} \quad (3.1.38)$$

which is minimized when ψ_{1+j}^* is set equal to ψ_{1+j} . This yields

$$E[z_{t+1} - \hat{z}_t(1)]^2 = (1 + \psi_{1+1}^2 + \dots + \psi_{1+l-1}^2) \sigma^2 \quad (3.1.39)$$

which turns out to be the 1 period ahead forecast error variance; by choosing the weights $\psi_j^* = \psi_j$, one implicitly sets $z_{t+1} = \hat{z}_t(1)$ and thereby chooses the unbiased forecast estimator.

In summary, the minimum mean square error forecast at time t for lead time 1 is the conditional expectation of z_{t+1} , given all information contained in the z 's up to and including time t . Additionally, forecasts from ARIMA models are optimal in the sense that no other linear univariate fixed coefficient estimator produces forecasts with smaller mean square error. Keep in mind, however, that these univariate forecasts are only optimal within their class and if the appropriate model has been found. Unfortunately, it is never clear whether the model chosen is the best in this sense.

3.1.3 Distributed Lag Models

There is no question that univariate models are useful when researchers lack information about factors affecting a time series. However, economists are often able to use economic theory to suggest possible determinants of the time series y_t . Specifically, the economist uses prior information of the form $E(y_t) = x_t' \beta$ where $x_t' = (x_{t1}, \dots, x_{tK})$, a $1 \times K$ row of independent variables at time t and β is a vector of unknown parameters. In time series analysis, the economic researcher may also know that past values of x_t' affect y_t , but is either unable or unwilling to specify how many lagged values of x_t' to include as regressors in a

statistical model.

The mean of the time series y_t , which depends on current and lagged values of K independent variables, can be expressed as

$$E[y_t] = \beta_1(L)x_{t1} + \beta_2(L)x_{t2} + \dots + \beta_K(L)x_{tK} \quad (3.1.40)$$

where $\beta_i(L)x_{ti} = \sum_{j=0}^{n(i)} \beta_{ij}x_{t-j}$, $i=1,2,\dots,K$, and $n(i)$ are unknown parameters. Let $\beta'(L) = [\beta_1(L), \beta_2(L), \dots, \beta_K(L)]$ and (3.1.12) can be written

$$\theta(L)[y_t - \beta'(L)x_t] = \alpha(L)e_t \quad (3.1.41)$$

Solving (3.1.41) for y_t yields

$$y_t = \beta'(L)x_t + [\alpha(L)/\theta(L)]e_t$$

or,

$$y_t = \beta'(L)x_t + \psi(L)e_t \quad (3.1.42)$$

where

$$\begin{aligned} \alpha(L) &= 1 + \alpha_1 L + \dots + \alpha_q L^q \\ \theta(L) &= 1 + \theta_1 L + \dots + \theta_p L^p \\ \beta_i(L) &= 1 + \beta_{i,1} L + \dots + \beta_{i,n(i)} L^{n(i)} \\ \psi(L) &= \alpha(L)/\theta(L) \\ e_t &\sim N(0, \sigma^2), \text{ and } \text{Cov}(e_t, e_s) = 0 \quad s \neq t. \end{aligned}$$

Notice that each explanatory variable is permitted to have a different lag length $n(i)$. If $n(i)$ is not finite for any i , then the model is said to have an infinite distributed lag. In this event, some kind of restrictions must be imposed on the parameters in order to obtain unique estimates using a finite sample. On the other hand, if

$\{n(i): i=1,2,\dots,K\}$ are finite, then the model is called a finite distributed lag model. As a special case, note that if $q=n(i)=0$ for all i , then equation (3.1.42) reduces to a simple linear regression model with autocorrelated errors. If each $n(i)=0$, then (3.1.42) is the ARMA(p,q) of the preceding section.

(a) Finite Distributed Lags of Unknown Length

Under the assumption that $p=q=0$ and that each $n(i)$ is finite, the model (3.1.42) becomes

$$y_t = \beta'(L)x_t + e_t \quad (3.1.43)$$

with $n=\max\{n(1),n(2),\dots,n(K)\}$, $t=n+1,\dots,T$. If each of the i^{th} lag lengths $(n(i) \ i=1,2,\dots,K)$ is known and if each of the e_t is i.i.d. normal, then maximum likelihood estimation of (3.1.43) yields m.v.u. estimates of β . However, $n(i)$ is seldom, if ever, known in economic research; consequently, estimation of equation (3.1.43) is problematic in several respects. First, the sample size is a function of an unknown parameter n and must be estimated. Second, from the discussion in section 2.3 it is clear that choosing any $n(i)$ too long leads to an inefficient estimator of $\beta'(L)$ and choosing any $n(i)$ too short will lead to a biased estimator of $\beta'(L)$. Third, pretesting to select $n(i)$ will be subject to the criticisms enumerated in section 2.4, namely the risk of using a pretest to find $n(i)$ will be greater than that of using the RLS estimator for all possible values of the true parameter vector and greater than that of using an overparameterized OLS

estimator over all but a small portion of the parameter space. In addition, the resulting estimators and statistics do not have the familiar central F or t-distributions. Nevertheless, most researchers continue to use preliminary tests to determine $n(i)$ before estimating $\beta(L)$. Because of the widespread use of these model selection techniques, and because the risk properties of the resulting estimators have yet to be studied, several means of specifying distributed lag models will be considered here.

(b) Estimating Lag Length

To simplify presentation of the various techniques used to estimate $n(i)$, it is assumed that y_t is systematically determined by current and lagged values of a single variable x and the intercept term is ignored. This model has the form

$$y_t = \sum_{i=0}^{n^*} \beta_i x_{t-i} + e_t \quad e_t \sim N(0, \sigma^2 I_{T-n^*-1}) \quad t=n^*+1, \dots, T \quad (3.1.44)$$

where n^* is the finite, but unknown lag length. The fact that n^* is unknown means that the number of lagged values of x to use as regressors is unknown. If too few regressors are included (the estimated value $n < n^*$) then the OLS estimator of β is biased and if too many are included ($n > n^*$) then the OLS estimator is inefficient. Many schemes have been proposed to aid researchers in selecting $n=n^*$. A traditional method selects an upper bound of n , say N , beyond which it is certain the β

coefficients are zero. Then, the researcher sequentially tests the relevance of the last $N-r$ coefficients

$$H_{(0)}: \beta_N = 0$$

$$H_{(1)}: \beta_N = \beta_{N-1} = 0$$

$$H_{(2)}: \beta_N = \beta_{N-1} = \beta_{N-2} = 0$$

.

$$H_{(r)}: \beta_N = \beta_{N-1} = \dots = \beta_{N-r} = 0$$

using the F-test. The criterion is to select the lag length based on the first hypothesis that is rejected. That is, if $H_{(r)}$ is rejected, $n=N-r$ is said to be the optimal lag length. Once n is selected, estimation proceeds as if the true model is the one under consideration. [Pagano and Hartley, (1973)] Note however, the resulting estimator is a pretest estimator; consequently, the unconditional sampling distribution of any of the usual statistics is unknown. Usual probability statements (e.g., t- and F-tests) made based on the estimated model must be conditioned on the validity of the model.

(c) Model Selection Criteria

Another approach uses model selection criteria to choose among the alternatives [Akaike (1974), Amemiya (1980), Mallows (1973), Parzen (1974)]. The model selection criteria operate on the following general principle. As the number of parameters included in a regression model increases, the calculated sample variance $\tilde{\sigma}_n^2$ declines. Thus, a penalty function is added to $\tilde{\sigma}_n^2$

which increases monotonically as regressors are added to the model.

Following Geweke and Meese (1981), suppose that the estimate of n^* , denoted n , is chosen to be the value which minimizes

$$E[C(n,T)] = \tilde{\sigma}_n^2 + ng(T) \quad n=0,1,\dots,N$$

where $C(n,T)$ is the criterion function, $\tilde{\sigma}_n^2 = \hat{e}'\hat{e}/T$, $g(T) > 0$ is a function of the sample size to be specified, and $N \leq T$ is the largest model to be considered. In most criterion functions the marginal penalty function $g(T)$ is proportional to $1/T$. Consequently, $g(T)$ becomes negligible asymptotically and the probability of underestimating n^* vanishes for large T . If the penalty function is too small, however, one tends to overestimate n^* on average.

Many criterion functions have been proposed. The following 3 criteria are useful when a choice must be made from many alternatives under the goal of risk minimization under a mean square error of prediction norm.

(i) Akaike's (1973) information criterion (AIC)

assumes the form

$$AIC(n,T) = \ln \tilde{\sigma}_n^2 + 2n/T$$

where $\tilde{\sigma}_n^2 = \hat{e}'\hat{e}/T$ evaluated under the assumption that $n=n^*$. The estimate of n is chosen such that

$$AIC(n,T) = \min\{AIC(n,T) \mid n=0,1,\dots,N\}$$

where N is the maximum lag length to be considered given that a sample of size T is available.

(ii) Amemiya's (1980) unconditional mean square error

criterion considers the risk of choosing an incorrect model under mean square error of prediction loss. Amemiya suggests minimizing

$$PC(n,T) = \tilde{\sigma}_n^2 [(T+n)/(T-n)].$$

(iii) Mallows (1973) suggests minimizing

$$Cp(n,T) = \tilde{\sigma}_n^2 + 2n\hat{\sigma}^2/T$$

where $\hat{\sigma}^2 = \hat{e}'\hat{e}/(T-K)$, the unbiased estimator of σ^2 .

Another approach is that taken by Schwarz (1978) who considers an infinite sequence of nested models, each of which has nonzero prior probability. When the sample distribution is normal, one selects the model with greatest posterior probability. Asymptotically, Schwarz shows that this is equivalent to minimizing

$$SBIC(n,T) = \ln \tilde{\sigma}_n^2 + n \ln(T)/T$$

which is referred to by Geweke and Meese (1981) as the Schwarz Bayesian Information Criterion. Geweke and Meese (1981) also consider a variant of the SBIC criterion which they call the Bayesian Estimation Criterion (BEC), denoted

$$BEC(n,T) = \tilde{\sigma}_n^2 + n \tilde{\sigma}_N^2 \ln(T)/(T-N).$$

Geweke and Meese [(1981), p. 57] show that under rather weak assumptions the probability of underestimating n^* vanishes for any of the above model selection criteria. On the other hand, for all of the criteria with the exception of SBIC and BEC, the probability of overestimating n^* does not vanish with large T . This implies that only SBIC and BEC lead to consistent estimation of the parameter n^* . By choosing n^* such that SBIC or BEC is minimized, one is

assured that the asymptotic distribution of the resulting estimator will be the same as if n^* were known. Though not consistent, the other criteria have been shown to have desirable properties also. For instance, Shibata (1981) proves that PC, AIC, and C_p choose the finite lag model that asymptotically minimizes sum of squared prediction errors.

(d) Estimating Lag Length for More than One Variable

Finally, if $i > 1$ i.e., there is more than one time series to be considered as an independent variable, one must select optimal lag lengths for each unknown parameter $n(i)$ $i=1,2,\dots,K$. In this case, one selects the specification which globally minimizes the model selection criterion $C(n,T)$, where n represents the total number of parameters estimated. The resulting equation will be of the form

$$y_t = \sum_{i=0}^{n(1)} \beta_{1,i} x_{1,t-i} + \sum_{i=0}^{n(2)} \beta_{2,i} x_{2,t-i} + \dots$$

$$+ \sum_{i=0}^{n(K)} \beta_{K,i} x_{K,t-i} + e_t$$

(3.1.52)

for $t=N+1,\dots,T$, and $K + \sum_{j=1}^K n(j) \leq T-N-1$.

There are at least two problems with estimation of (3.1.52): (1) It is often likely that the number of parameters to be estimated $K + \sum n(j)$ will be large relative to the number of available observations $T-N-1$. (2)

Multicollinearity is often quite severe when several lagged values of an independent variable are included as regressors and precise estimation of one or more of the β_{ij} may be difficult. One technique used to mitigate both of these problems is the Almon Lag procedure. Almon's (1965) solution to these two problems was to force the lag weights β_{ij} to fall along a polynomial of degree $r(i) \leq n(i)$ for $i=1,2,\dots,K$. In the section which follows, a variation of the Almon procedure will be discussed which permits the use of the RLS framework of Section 2.4. Then, a few comments will be made summarizing the consequences of incorrectly specifying lag length and polynomial degree.

3.1.4 Polynomial Distributed Lag Models

Consider again the model (3.1.44), and assume that the lag length n^* has been determined to equal n .

$$y_t = \sum_{i=0}^n \beta_i x_{t-i} + e_t \quad e_t \sim N(0, \sigma^2 I) \quad t=n+1, \dots, T \quad (3.1.53)$$

Suppose that β_j can be expressed as a polynomial of order r in the integers $j=0,1,\dots,n$. That is to say

$$\beta_j = \alpha_0 + \alpha_1 j + \alpha_2 j^2 + \dots + \alpha_r j^r \quad (3.1.54)$$

for $j=0,1,\dots,n$ and $r \leq n$ where the α_i are unknown parameters. Notice that the $n+1$ β 's are now expressed as functions of $q+1$ α 's. This reparameterization has reduced the number of parameters to be estimated by imposing $n-q$ exact restrictions on the β_j , and implies that the dimension of the estimation problem can be reduced by imposing polynomial restrictions on the shape of the

distributed lag. In fact, it is possible to express these restrictions as a set of linear homogeneous equations, thus permitting the use of the RLS estimator (2.4.2) and the Stein-rule estimator (2.6.2).

(a) Polynomial Restrictions

Several methods have been suggested for specifying the matrix R needed for use in the RLS estimator (2.4.2) [see Hill (1982)]. The method developed here was devised by Hill and offers significant computational advantages over equivalent methods. Hill recommends using orthogonal polynomials to derive the restriction matrix; the method is computationally efficient because R need not be recomputed as a nested set of polynomial restrictions is sequentially tested. This feature is especially attractive if the PDL model is to be used in Monte Carlo experiments.

Consider the polynomial equation

$$H_r(i) = \alpha_{0,r} + \alpha_{1,r}i + \alpha_{2,r}i^2 + \dots + \alpha_{r,r}i^r \quad (3.1.55)$$

where $i=0,1,2,\dots,n$. These polynomials can be constructed in such a way that

$$\sum_{i=0}^n H_j(i) H_m(i) = 0 \quad j \neq m$$

and are therefore orthogonal. The β_j can be expressed exactly as functions of these orthogonal polynomials

$$\beta_j = d_0 H_0(i) + d_1 H_1(i) + \dots + d_q H_q(i) \quad j=1,2,\dots,n$$

or in matrix notation

$$\beta = Hd. \quad (3.1.56)$$

Pre-multiplying (3.1.56) by $(H'H)^{-1}H'$ yields

$$d = (H'H)^{-1}H'\beta. \quad (3.1.57)$$

Because the matrix H is orthogonal, $H'H$ is diagonal and the

j^{th} diagonal element can be expressed as $A_j = \sum_{i=0}^n \{H_j(i)\}^2$.

Using this fact, the j^{th} element of the vector a can be denoted

$$d_j = \sum_{i=0}^n \beta_i H_j(i) / A_j \quad j=0,1,2,\dots,r.$$

For all $j > r$, d_j will be zero, implying a set of homogeneous equations that can serve as restrictions for the RLS estimator. Specifically, to estimate a polynomial of degree $q < n$, one sets $d_{q+1} = \dots = d_n = 0$. This is done by deleting $n-q$ rows of the $(n+1) \times (n+1)$ matrix of orthogonal polynomials. Tabled values of the $H_j(\blacksquare)$ are available [Delury (1950)] or they can be generated using SAS or some other computer software package.

(b) Model Selection

The strategy for estimating a polynomial of degree r is much the same as that of estimating lag length. The degree of polynomial r can be chosen either by minimization of a model selection criterion or by sequentially testing the nested restrictions

$$\begin{aligned}
 H_n &: \sum_{i=0}^n H_n(i) \beta_i = 0 \\
 H_{n-1} &: \sum_{i=0}^n H_n(i) \beta_i = 0 \\
 &\quad \sum_{i=0}^n H_{n-1}(i) \beta_i = 0 \\
 &\quad \cdot \\
 &\quad \cdot \\
 &\quad \cdot \\
 H_r &: \sum_{i=0}^n H_n(i) \beta_i = 0 \\
 &\quad \sum_{i=0}^n H_{n-1}(i) \beta_i = 0 \\
 &\quad \cdot \\
 &\quad \cdot \\
 &\quad \cdot \\
 &\quad \sum_{i=0}^n H_r(i) \beta_i = 0
 \end{aligned} \tag{3.1.58}$$

using the F-test. For models with more than one explanatory variable, one may proceed in much the same way as in selecting lag lengths for more than one set of explanatory variables. The lag lengths $n(i)$ are assumed to be the true ones and optimal polynomial degrees are chosen such that the criterion function $C(n, T)$ is minimized over

all possible polynomial combinations.

(c) Effects of Misspecification

Finally, mention should be made of the ill-effects of misspecifying lag lengths or polynomial degrees. These effects, presented in Trivedi and Pagan (1979) and neatly summarized by Judge et al. [(1985), pp. 359-360] are essentially reproduced below.

1. If the assumed polynomial degree is correct, but the lag length too long, the PDL estimator will generally be biased. If the difference between the true lag length and the estimated one is greater than the degree of the polynomial, then the estimator is always biased.

2. If the estimated polynomial degree is correct, then underestimating $n(i)$ results in a biased estimator.

3. If the estimated lag length is correct, but the estimated polynomial is too high, the PDL estimator is unbiased and inefficient.

4. If the estimated lag length is correct but the degree of the estimated polynomial degree is too low, then the PDL estimator is biased.

Unlike the case where only lag lengths are to be estimated, researchers cannot depend on techniques which asymptotically overestimate $n(i)$. Once polynomial restrictions are added to the lag restrictions one runs the risk of overstating $n(i)$ and understating the polynomial. Such an event will "stretch" the lag effects to regions of the parameter space where they do not belong and result in

a biased estimator. Furthermore, it seems rather imprudent to place much faith in hypothesis tests based on the resulting model. The model specification process in effect discards sample information inconsistent with the model itself. As a result, the computed standard errors are likely to understate the true ones. One additional warning is required. Many estimators which are preferred on the basis of their large sample properties may actually perform worse than competitors in the small samples typically encountered in econometric practice.

In some cases, one is not concerned with the resulting sampling properties of models specified on the basis of the sample information. If the goal is to build a forecasting model of the dependent variable, then model selection procedures can often be very useful. Assessing out-of-sample forecast accuracy of competing models and estimators is important in several respects. Economists engaged in business or government are often called upon to generate forecasts and seldom "have time" to engage in sophisticated model selection processes. A desirable estimator for these practitioners is one which is easy to compute and yields good forecasts. The Stein-rule estimator is certainly a candidate since it obviates the need for a specification search and it is easy to compute. Whether or not it yields good forecasts in realistic situations remains to be seen. In order to assess the forecast accuracy of various estimators in Chapter 4, it is perhaps useful here to

describe several of the most frequently used measures of gauging forecast accuracy.

(d) Forecast Evaluation

In Chapter 4 the out-of-sample predictive accuracy of several Stein-rule estimators will be compared to traditional rivals OLS, RLS (PDL/Pretest), and univariate ARIMA. The three most widely used measures of forecast accuracy are root mean square error (RMSE), mean absolute error (MAE), and Theil's inequality coefficient (U). The three measures have recently been described by Fair (1986) and are discussed below.

First, let \hat{y}_t be the forecasted value of y for time period t and let y_t be the actual value of the time series. Assuming that l post sample observations on y are available, the forecast error can be evaluated using any one of the following measures:

$$\text{RMSE} = [l^{-1}(\mathbf{y}'\hat{\mathbf{y}})]^{-\frac{1}{2}} \quad (3.1.59)$$

$$\text{MAE} = l^{-1} \sum_{t=1}^l |y_t - \hat{y}_t| \quad (3.1.60)$$

$$U = [l^{-1}\Delta y_t' \Delta \hat{y}_t]^{-\frac{1}{2}} / [l^{-1}\Delta y_t' \Delta y_t]^{-\frac{1}{2}} \quad (3.1.61)$$

where Δ in (3.1.61) denotes either the absolute or percentage change.

If forecasts are perfect, all three measures will be zero. The RMSE criterion, which is a function of the squared deviations of the forecasts from actual values of y , penalizes large errors to a greater extent than MAE.

Theil's inequality coefficient is most useful in those circumstances when no competing model is being considered. Notice that $\Delta \hat{y}_t = 0$ implies that there is no change in forecasted values of y_t . In this instance, $U=1$. If $U>1$, then the forecast error is greater than it would have been if a no-change forecast had been made; the forecasting equation fails in the sense that naively forecasting \hat{y}_{t+1} to be the same as y_t would have been a closer approximation to what actually happened.

In the next chapter only the RMSE measure will be computed since interest lies in comparing alternative predictive models. As a measure of forecast error, RMSE is also consistent with the use of quadratic risk functions as a basis for judging estimator performance.

3.2 Nonlinear Models

When economic theory fails to suggest a specific functional form, empirical researchers are often willing to assume that the approximate relationship among the variables is linear, or that the relationship is linear after an appropriate transformation of the variables [see Spitzer, (1982a), (1982b)]. Although the parameters of a linear equation like (2.1.1) are easy to estimate, the linear assumption underlying the model can seldom be justified on the basis of economic theory. If the true model is not linear, or at least approximately so, then the least squares estimator will generally be biased and inconsistent.

Many econometric models are inherently nonlinear in the parameters. For instance, the parameters of the CES production function cannot be transformed and estimated using traditional linear techniques like OLS. With the growing interest in models of this type, it is important for empirical economic researchers to acquire a working knowledge of nonlinear estimation techniques.

Given the apparent necessity of nonlinear estimation in econometrics and the recent interest in certain improved methods of point estimation (Stein-rules, empirical Bayes, etc.), it is also important to ask whether improved estimators of the parameters of nonlinear models can be found. To date, a Stein-rule estimator has not been formulated for this important class of models.

The search for an improved estimator of the parameters of a nonlinear statistical model begins with the probit regression. As Finney (1952) has demonstrated, the maximum likelihood parameter estimates of a probit model can be obtained using iterated weighted least squares. Nelder and Wedderburn (1972) generalize Finney's result for a much wider class of models, the so-called generalized linear models (GLIM). The least squares interpretation of the GLIM estimates suggests an algorithm for developing a Stein-like shrinkage estimator of the nonlinear probit regression model.

In order to lay a proper foundation for an investigation of a Stein-like shrinkage estimator of the parameters of a probit regression model, several issues will be discussed. In the remainder of this section key results from the theory of nonlinear statistical models will be briefly summarized. Then, the rudiments of numerical optimization techniques will be presented. In section 3.3 a rather detailed exposition of Nelder and Wedderburn's (1972) generalized linear model will be given. This is followed by the example of interest, the probit regression model. In the final section of this chapter, two computer intensive research techniques--Monte Carlo and bootstrapping--will be introduced. These two methods are to be used to evaluate the unknown sampling properties of various Stein-rule estimators used in Chapters 5 and 6.

3.2.1 Nonlinear Statistical Model

Consider the following nonlinear statistical model

$$y = f(\beta, X) + e \quad (3.2.1)$$

where y is a $T \times 1$ vector of observable random variables, f is a $T \times 1$ vector valued function, β is a $K \times 1$ vector of unknown parameters, X is a $T \times M$ nonstochastic matrix of treatment variables, and e is a $T \times 1$ vector of independent unobservable random variables, each with zero mean and variance σ^2 . For notational simplicity, equation (3.2.1) is sometimes written as

$$y = f(\beta) + e,$$

a convention which shall be followed below.

The nonlinear least squares estimator (NLLS) is defined to be the value of the vector β which minimizes the sum of squared errors function

$$s(\beta) = \sum_1^T [y_t - f_t(\beta)]^2 \quad (3.2.2)$$

over the parameter space B . Minimization of the sum of squares function with respect to β yields

$$g(\beta) = \partial s / \partial \beta = -2 \sum_1^T (\partial f_t / \partial \beta) [y_t - f_t(\beta)] \quad (3.2.3)$$

where $\partial f_t / \partial \beta = [\partial f_t / \partial \beta_1, \dots, \partial f_t / \partial \beta_K]'$. Second order conditions require the Hessian matrix

$$\begin{aligned} H(\beta) = \partial^2 s / \partial \beta \partial \beta' = & -2 \left[\sum_1^T (\partial^2 f_t / \partial \beta \partial \beta') [y_t - f_t(\beta)] \right. \\ & \left. + \sum_1^T (\partial f_t / \partial \beta) (\partial f_t / \partial \beta)' \right], \end{aligned} \quad (3.2.4)$$

to be positive definite, where

$$\partial^2 f_t / \partial \beta \partial \beta' = \begin{bmatrix} \partial^2 f_t / \partial \beta_1 \partial \beta_1 & \partial^2 f_t / \partial \beta_1 \partial \beta_2 & \dots & \partial^2 f_t / \partial \beta_1 \partial \beta_K \\ \partial^2 f_t / \partial \beta_2 \partial \beta_1 & \partial^2 f_t / \partial \beta_2 \partial \beta_2 & \dots & \partial^2 f_t / \partial \beta_2 \partial \beta_K \\ \vdots & \vdots & \ddots & \vdots \\ \partial^2 f_t / \partial \beta_K \partial \beta_1 & \partial^2 f_t / \partial \beta_K \partial \beta_2 & \dots & \partial^2 f_t / \partial \beta_K \partial \beta_K \end{bmatrix}.$$

Unlike the linear regression model (2.1.1), the solutions to the normal equations (3.2.3) cannot be obtained analytically. Consequently, numerical techniques like those discussed in the next section must be used to derive the minimizing vector b .

Given the NLLS estimator b of β , define the NLLS estimator of σ^2 to be $\hat{\sigma}^2 = S_T(b)/T$. If the e_t are assumed to be independent and normally distributed, then the maximum likelihood estimators of β and σ^2 are also b and $\hat{\sigma}^2$, respectively.

(a) Statistical Properties

Several additional assumptions are required to ensure that NLLS yields consistent estimates of β . Amemiya (1985, p. 129) provides the following sufficient conditions for consistency:

1. $\partial f / \partial \beta$ exists and is continuous.
2. The parameter space \mathbf{B} is compact.
3. $T^{-1}f(\beta_1)'f(\beta_2)$ converges uniformly in $\beta_1, \beta_2 \in \mathbf{B}$.
4. $\lim T^{-1}[f(\beta^*) - f(\beta)]'[f(\beta^*) - f(\beta)]$ exists and is nonzero if $\beta \neq \beta^*$.

As Amemiya points out, these assumptions have rough counterparts in the linear model (2.1.1). Furthermore, they will generally be met for the regular concave density functions like the ones considered below and therefore will not be discussed further.

In addition, Amemiya [(1985), Theorem 4.3.2, pp. 133-134] proves the following result. If

i. $\lim T^{-1} G(\beta^*)'G(\beta^*) \equiv C$ is a finite nonsingular matrix,

ii. $T^{-1} G(\beta)'G(\beta)$ converges to a finite matrix uniformly for all β in an open neighborhood of β^* ,

iii. $\partial^2 f_t / \partial \beta_i \partial \beta_j$ is continuous in β in an open neighborhood of β^* uniformly in t , $i, j=1, 2, \dots, K$,

iv. $\lim T^{-2} \sum_1^T [\partial^2 f_t / \partial \beta_i \partial \beta_j]^2 = 0$ for all β in an open

neighborhood of β^* , and

v. $T^{-1} \sum_1^T f_t(\beta_1) [\partial^2 f_t / \partial \beta \partial \beta']|_{\beta_2}$ converges to a finite

matrix uniformly for all β_1 and β_2 in an open neighborhood of β^* , then the NLLS estimator has the following limiting distribution

$$T^{1/2} (b - \beta) \sim N(0, \sigma^{*2} [\lim G(\beta)'G(\beta)/T]^{-1}) \quad (3.2.5)$$

where $\sigma^{*2} = \text{plim } e'e/T$ and

$$G(\beta)' \equiv \partial f' / \partial \beta = \begin{bmatrix} \partial f_1 / \partial \beta_1 & \partial f_2 / \partial \beta_1 & \dots & \partial f_T / \partial \beta_1 \\ \partial f_1 / \partial \beta_2 & \partial f_2 / \partial \beta_2 & \dots & \partial f_T / \partial \beta_2 \\ \vdots & \vdots & \ddots & \vdots \\ \partial f_1 / \partial \beta_K & \partial f_2 / \partial \beta_K & \dots & \partial f_T / \partial \beta_K \end{bmatrix}.$$

The covariance matrix is estimated using

$$\hat{\sigma}^2 [G(b)'G(b)/T]^{-1}, \quad (3.2.6)$$

where $G(b)' = G(\beta)'I_b$, and $\hat{\sigma}^2 = s(b)/T$ or $s(b)/(T-K)$ (since they are asymptotically equivalent).

(b) Hypothesis Testing

Asymptotically, $b - \beta \cong [G(\beta)'G(\beta)]^{-1}G(\beta)'e$. Using the estimated value of $G(\beta) = G(b)$, it is possible to generalize the usual F-statistic for testing joint hypotheses, provided T is sufficiently large. One way to do this is to set $J \leq K$ of the elements of the vector β (denoted β_J) to their hypothesised values β_0 . The sum of squares function is minimized with respect to β subject to $\beta_J = \beta_0$, and denoted $s(\beta_0)$. A test of the null hypothesis $H_0: \beta_J = \beta_0$ against the alternative $H_a: \beta_J \neq \beta_0$ can be performed by evaluating the statistic

$$(T-K) [s(\beta_0) - s(b)] / Js(b) \quad (3.2.7)$$

which has an $F_{J, T-K}$ asymptotic distribution under the null hypothesis. Alternative means of testing hypotheses may be employed. Gallant (1975) has compared the small sample properties of similar test statistics in a Monte Carlo study and concludes that (3.2.7) tends to perform better than its competitors.

3.2.3 Numerical Optimization

In general, nonlinear estimation problems are solved through iterative procedures which directly minimize an objective function. Let the objective function be denoted $Q(\theta)$, where θ is a $K \times 1$ vector of unknown parameters. In principle, an initial parameter estimate θ_0 or starting

value is obtained; based on this estimate a new "improved" estimate θ_1 is computed. The new estimate is then used to produce another improvement manifested in θ_2 . This process continues until no further significant improvements can be obtained, i.e., until convergence.

There are a number of competing algorithms available for use in numerical optimization. They differ to the extent that they employ mathematical expectations and partial derivatives of the objective function, (i.e., the sum of squares function or the likelihood function). Some require only first partial derivatives [Berndt, Hall, Hall, Hausman (1974)], some require first and second partial derivatives (Newton-Raphson), and some require the expectation of the Hessian (method of scoring).

The purpose of numerical optimization is to minimize the objective function $Q(\theta)$ with respect to the K elements of the parameter vector θ . The gradient vector and the Hessian matrix are defined to be

$$g(\theta) = \partial Q(\theta) / \partial \theta \quad (\text{gradient})$$

$$H(\theta) = \partial^2 Q(\theta) / \partial \theta \partial \theta'. \quad (\text{Hessian})$$

If $g(\theta)$ and $H(\theta)$ exist and are continuous in the neighborhood of a particular value $\hat{\theta}$, sufficient conditions for a local minimum of $Q(\theta)$ at $\hat{\theta}$ are that $g(\hat{\theta}) = 0$ and $H(\hat{\theta})$ be positive definite.

Each of the numerical optimization techniques considered below takes the form

$$\theta_{n+1} = \theta_n + \lambda_n C(\theta_n) d(\theta_n)$$

where θ_n is the n^{th} round estimate of θ ,

$C(\theta_n)$ is a positive definite matrix,

λ_n is a positive scalar called the step length, and

$d(\theta_n)$ is a direction vector which is a function of the gradient. The choice of λ_n and $d(\theta_n)$ will be discussed shortly. First, it is important to define exactly what is meant by convergence.

Definition 3.2.1 An iterative scheme is said to have converged when one or more of the following conditions are met:

1. $Q(\theta_{n+1})$ is "close" to $Q(\theta_n)$ in the sense that
 $|Q(\theta_{n+1}) - Q(\theta_n)| < \epsilon_1, \epsilon_1 > 0$
2. θ_{n+1} is "close" to θ_n in the sense that
 $(\theta_{n+1} - \theta_n)'(\theta_{n+1} - \theta_n) < \epsilon_2, \epsilon_2 > 0$
3. $g(\theta_{n+1})$ is "close" to $g(\theta_n)$ in the sense that
 $[g(\theta_{n+1})]'[g(\theta_{n+1})] < \epsilon_3, \epsilon_3 > 0.$

In practice, all three conditions should be verified for sufficiently small values of ϵ_1 , ϵ_2 , and ϵ_3 . Furthermore, if the objective function is not strictly concave (or convex), then several different starting values θ_0 should be tried to ensure that the objective function is globally minimized (maximized).

(a) Newton-Raphson

The Newton-Raphson algorithm is based on a second order Taylor series expansion of the objective function $Q(\theta)$ around the estimated value of the parameter vector $\hat{\theta}$.

Expansion of $Q(\theta)$ yields

$$Q(\theta) \cong Q(\hat{\theta}) + g(\hat{\theta})'(\theta - \hat{\theta}) + 1/2 (\theta - \hat{\theta})' H(\hat{\theta}) (\theta - \hat{\theta}).$$

where $\hat{\theta}$ is an initial estimate of θ , $g(\hat{\theta})' = [\partial Q / \partial \theta']|_{\hat{\theta}}$, and $H(\hat{\theta}) = [\partial^2 Q / \partial \theta \partial \theta']|_{\hat{\theta}}$. Differentiation of the right-hand-side with respect to θ yields

$$g(\theta) \cong \partial Q(\theta) / \partial \theta = g(\hat{\theta}) + H(\hat{\theta})(\theta - \hat{\theta}).$$

Using the fact that $g(\theta) = 0$ when the objective function is minimized, one can solve for θ :

$$\begin{aligned} [H(\hat{\theta})]\theta &\cong [H(\hat{\theta})]\hat{\theta} - g(\hat{\theta}) \\ \theta &\cong \hat{\theta} - H^{-1}(\hat{\theta})g(\hat{\theta}). \end{aligned} \quad (3.2.8)$$

The equality in (3.2.8) is exact if $Q(\theta)$ is quadratic; in this case, $\hat{\theta}$ can be set to any initial estimate and the Newton-Raphson will converge to a global minimum in one iteration. If $Q(\theta)$ is of an order higher than 2, then the use of (3.2.8) iteratively may lead to convergence at a local minimum. If H^{-1} is not positive definite, then the Newton-Raphson searches for the minimum of the objective function in the wrong direction. That is, (3.2.8) may locate a local or global maximum. Many solutions to these problems have been proposed. Several of these (method of scoring, quadratic hill climbing, method of steepest ascent) are mentioned below.

Finally, the Newton-Raphson algorithm as given in (3.2.8) implicitly chooses the step length to be 1, i.e., $\lambda_n = 1$. In general, one would prefer to take larger steps when far away from an extreme value and smaller steps when close. If the step size is too large, then one can

overshoot the target value and if the step length is too small, convergence may take an inordinate number of iterations. Various methods of selecting an appropriate step size have been mentioned in the literature [Goldfeld et al. (1966), Goldfeld and Quandt (1972)]. The method of steepest ascent [Fletcher and Powell (1963)] addresses the problem of varying the step length by choosing λ_n analytically as a function of the Hessian and the gradient. Quadratic hill climbing [Goldfeld et al. (1966)] uses a measure of the accuracy of the quadratic approximation to choose an appropriate step length. For a summary of these and other techniques, consult Bard (1974) or Goldfeld and Quandt (1966).

In summary, given proper choices of step length and starting values, the following equation used iteratively will yield parameter estimates which locally minimize the objective function $Q(\theta)$:

$$\theta_{n+1} = \theta_n - \lambda_n H^{-1}(\theta_n) g(\theta_n). \quad (3.2.9)$$

(b) Method of Scoring

If the mathematical expectation of the Hessian matrix is easy to take, then the resulting Information matrix may be used in its place, i.e.,

$$C(\theta_n) = -E[H(\theta_n)] = I(\theta)$$

where $I(\theta)$ is Fisher's Information matrix (assuming that $Q(\theta)$ is the log likelihood function). Unlike the Newton-Raphson, the method of scoring guarantees that $C(\theta_n)$ will be positive definite; therefore, the numerical search is

conducted in an appropriate direction (relative to the starting values). And as usual, it is advisable to use a variable step length λ_n . Thus, the method of scoring yields the following iterative estimator of the unknown parameter vector θ

$$\theta_{n+1} = \theta_n + [I(\theta_n)]^{-1}g(\theta_n) \quad (3.2.10)$$

If $I(\theta)$ is block diagonal, then this procedure may simplify computation by reducing the number of elements to be estimated.

(c) Gauss-Newton

Recall that in order to derive NLLS or MLE estimates of the parameters of the nonlinear regression model, the sum of squares function $s(\beta)$ must be minimized. The Gauss-Newton algorithm is derived by replacing $f(\beta)$ in the sum of squares function with its first order Taylor's series approximation taken at an estimate b . Thus, the sum of squares function is

$$s(\beta) = [y - f(\beta)]' [y - f(\beta)]. \quad (3.2.11)$$

Taking the first order Taylor's series approximation of $f(\beta)$ about the estimate b_0 yields

$$f(\beta) \cong f(b_0) + G(b_0)(b_0 - \beta) \quad (3.2.12)$$

where

$$G(\beta)' \equiv \partial f' / \partial \beta = \begin{bmatrix} \partial f_1 / \partial \beta_1 & \partial f_2 / \partial \beta_1 & \dots & \partial f_T / \partial \beta_1 \\ \partial f_1 / \partial \beta_2 & \partial f_2 / \partial \beta_2 & \dots & \partial f_T / \partial \beta_2 \\ \vdots & \vdots & \ddots & \vdots \\ \partial f_1 / \partial \beta_K & \partial f_2 / \partial \beta_K & \dots & \partial f_T / \partial \beta_K \end{bmatrix}$$

and $G(b_0) = G(\beta)|_{b_0}$. Substituting (3.2.12) into the sum of

squares function (3.2.11) yields

$$s(\beta) \equiv \{[y - [f(b_0) + G(b_0)(b_0 - \beta)]]'\{y - [f(b_0) + G(b_0)(b_0 - \beta)]\}\}.$$

Minimization of $s(\beta)$ with respect to β yields

$$2G(b_0)'y - 2G(b_0)'f(b_0) + 2G(b_0)'G(b_0)(b_0 - \beta).$$

Denoting the minimizing value of β as b_1 ,

$$2G(b_0)'y - 2G(b_0)'f(b_0) + 2G(b_0)'G(b_0)b_0 - 2G(b_0)'G(b_0)b_1 = 0$$

dividing both sides of the equation by 2,

$$G(b_0)'G(b_0)b_1 = G(b_0)'y - G(b_0)'f(b_0) + G(b_0)'G(b_0)b_0$$

and rearranging yields

$$G(b_0)'G(b_0)b_1 = G(b_0)'G(b_0)b_0 + G(b_0)'[y - f(b_0)].$$

Finally, pre-multiplying both sides of the equality by

$[G(b_0)'G(b_0)]^{-1}$ yields

$$b_1 = b_0 + [G(b_0)'G(b_0)]^{-1}G(b_0)'[y - f(b_0)]$$

or,

$$b_1 = [G(b_0)'G(b_0)]^{-1}G(b_0)'[y - f(b_0) + G(b_0)b_0].$$

(3.2.13)

The new estimate b_1 is then substituted for the initial estimate b_0 in the right-hand-side of (3.2.13) to produce a second round estimate b_2 . This process is repeated until the iterative scheme converges. Notice that equation (3.2.13) has a similar form to the least squares estimator $b = (X'X)^{-1}X'y$, where X is the gradient matrix $G(b_0)$ and y is the adjusted independent variable $y - f(b_0) + G(b_0)b_0$. Given this interpretation, the $(n+1)^{th}$ round estimate b_{n+1} can be thought of as the least squares estimator of β from the model

$$\bar{y}(b_n) = G(b_n)\beta + e \quad (3.2.14)$$

where $\bar{y}(b_n) = y - f(b_n) + G(b_n)b_n$ is an adjusted dependent variable and $G(b_n)$ is the "regressor" matrix. Both are functions of the current round estimate b_n and as a result, are expected to change with each successive iteration (until convergence).

These numerical optimization algorithms are of particular interest in this dissertation because of their use in generalized linear models (GLIM) of Nelder and Wedderburn (1972). According to Copas (1983), the idea of shrinkage is not confined to linear regression, but also applies to the much wider class of generalized linear models. In fact he considers briefly the binary regression model as a case in point. In Chapter 6, maximum likelihood estimates of the parameters of a probit regression model are obtained and these estimates are then used in a Stein-like shrinkage estimator of the probit model. To facilitate this end, the probit model is described below.

3.2.3 Example: The Probit Regression Model

To introduce the probit regression model, consider the following definition.

Definition 3.2.2 Qualitative response models are regression models in which the dependent variable takes discrete values. When a single dependent variable takes the value of 1 or zero the model is called the binary choice model and is defined to be

$$\Pr(y_t=1) = F(x_t'\beta) \quad t=1,2,\dots,T \quad (3.2.15)$$

where $\{y_t\}$ is a sequence of independent binary random variables taking the value 1 or zero, x_t is a known $K \times 1$ vector of explanatory variables associated with the t^{th} observation, β is a $K \times 1$ vector of unknown parameters, and F is a certain cumulative distribution function (c.d.f.).

Choosing F to be a c.d.f. ensures that $x_t'\beta$ is mapped onto the interval $[0,1]$.

The two most common choices of F are the normal c.d.f. and the logistic c.d.f.

$$\text{(Normal)} \quad F(x_t'\beta) = \Phi(x_t'\beta) = \int_{-\infty}^{x_t'\beta} (2\pi)^{-\frac{1}{2}} \exp\{-\frac{1}{2}r^2\} dr \quad (3.2.16)$$

$$\text{(Logistic)} \quad F(x_t'\beta) = \Lambda(x_t'\beta) = e^{x_t'\beta} / (1 + e^{x_t'\beta}) \quad (3.2.17)$$

The probit and logit functions are the inverses of the normal and logistic c.d.f.'s respectively (see equations 3.3.5 and 3.3.6 below). Although the two functions are similar in many respects, the choice between the logit or probit function is not an arbitrary one; the model selected should be that which is most consistent with the underlying data generation process.

The use of the probit function is often preferred and is motivated in the following way [see Fomby, Hill, Johnson (1986), p. 344]. Let A be the action taken if the expected utility associated A is great enough. If action A is taken, it is assigned the number 1; if not, it is assigned the number 0. Whether or not the action is taken depends on the value of the expected utility index I associated

with the consumption of a given bundle of goods and services X . The index variable I may be thought of econometrically as a latent variable determined linearly by the observable attributes X . Thus, $I_t = x_t'\beta$ is considered to be positively related to the probability that action A will be taken; the higher the index I , the more likely A .

Each person either takes or does not take action A based on whether the observed level of the index I is above or below his personal threshold level I_t^* . Each person's threshold level is a function of individual tastes and preferences; and, by the central limit theorem, tastes and preferences will be normally distributed across the population at large. Thus,

$$\pi_t = \Pr[A|I_t] = \Pr[I_t^* \leq I_t] = \Phi(x_t'\beta) \quad (3.2.18)$$

where again, Φ is the standard normal c.d.f. evaluated at the observed level of the index $x_t'\beta$.

Given that the probit model is consistent with the underlying data generation process, the estimation of the parameters of the probit regression model can be developed in light of the preceding discussion of nonlinear estimation techniques.

Let $\{y_1, y_2, \dots, y_T\}$ be a random sample of T Bernoulli trials with parameter π . The probability density function is denoted

$$f(y_t|\pi) = \pi^{y_t} (1-\pi)^{1-y_t} \quad t=1,2,\dots,T$$

Now suppose that (definition 3.2.1)

$$\pi_t = \Pr(y_t=1) = F(x_t'\beta) \quad t=1,2,\dots,T$$

and that the normal c.d.f. is the appropriate model of the parameter π . Thus,

$$\pi_t = F(x_t'\beta) = \Phi(x_t'\beta) \quad t=1,2,\dots,T.$$

The likelihood function is denoted

$$L = \prod_{t=1}^T \Phi(x_t'\beta)^{y_t} [1-\Phi(x_t'\beta)]^{1-y_t} \quad (3.2.19)$$

and the log-likelihood is

$$\ell = \ln L = \sum_{t=1}^T y_t \ln(\Phi_t) + (1-y_t) \ln(1-\Phi_t) \quad (3.2.20)$$

where $\Phi_t = \Phi(x_t'\beta)$.

The Newton-Raphson method requires the gradient vector $\partial \ell / \partial \beta$ and the Hessian matrix $\partial^2 \ell / \partial \beta \partial \beta'$. The gradient vector is

$$\partial \ell / \partial \beta = \sum_{t=1}^T y_t [\phi_t / \Phi_t] x_t - (1-y_t) [\phi_t / (1-\Phi_t)] x_t \quad (3.2.21)$$

where ϕ_t is the standard normal p.d.f. evaluated at the argument $x_t'\beta$. Equation (3.2.21) can be simplified to:

$$g(\beta) = \partial \ell / \partial \beta = \sum_{t=1}^T \{(y_t - \Phi_t) / [\Phi_t(1-\Phi_t)]\} \phi_t x_t \quad (3.2.22)$$

The Hessian is

$$H(\beta) = - \sum_{t=1}^T \phi_t \{ y_t [(x_t'\beta) \Phi_t + \phi_t] / (\Phi_t)^2 + (1-y_t) [\phi_t - (x_t'\beta)(1-\Phi_t)] / (1-\Phi_t)^2 \} x_t x_t'. \quad (3.2.23)$$

The global concavity of the likelihood function permits a liberal choice of starting values; those from the OLS estimator $b_0 = (X'X)^{-1}X'y$ are often used as the first approximation. Evaluating (3.2.22) and (3.2.23) at the

starting value b_0 and substituting into Newton-Raphson equation (3.2.9) with step length 1 yields

$$b_1 = b_0 - H^{-1}(b_0)g(b_0) \quad (3.2.24)$$

which is used iteratively until convergence.

The method of scoring technique can be used in a similar way; the only difference is that the negative of the Hessian is replaced with the Information matrix. For regular densities,

$$I(\beta) = -E[\partial^2 \ell / \partial \beta \partial \beta'] = E[(\partial \ell / \partial \beta)(\partial \ell / \partial \beta)'].$$

In terms of equation (3.2.22)

$$\begin{aligned} I(\beta) &= E[(\partial \ell / \partial \beta)(\partial \ell / \partial \beta)'] = E[g(\beta)g(\beta)'] \\ &= E \sum_1^T \{(y_t - \pi_t)^2 / [\pi_t(1 - \pi_t)]^2\} \phi_t^2 x_t x_t'. \end{aligned} \quad (3.2.25)$$

If the random variable is Bernoulli distributed [i.e., $y \sim b(1, \pi)$], and π is assumed to equal π_t , then

$$\begin{aligned} E(y_t) &= \pi_t = \pi_t \\ \text{Var}(y_t) &= \pi_t(1 - \pi_t) = \pi_t(1 - \pi_t). \end{aligned}$$

Using these facts, (3.2.25) may be written as

$$\begin{aligned} I(\beta) &= \sum_1^T \{E(y_t - \pi_t)^2 / [\pi_t(1 - \pi_t)]^2\} \phi_t^2 x_t x_t' \\ &= \sum_1^T \{\text{Var}(y_t) / [\pi_t(1 - \pi_t)]^2\} \phi_t^2 x_t x_t' \\ &= \sum_1^T \{\pi_t(1 - \pi_t) / [\pi_t(1 - \pi_t)]^2\} \phi_t^2 x_t x_t' \\ I(\beta) &= \sum_1^T \{[\phi_t^2 / [\pi_t(1 - \pi_t)]]\} x_t x_t' \end{aligned} \quad (3.2.26)$$

Although the Newton-Raphson and the method of scoring lead to the same point estimates in the probit regression model (due to the global concavity of Q), each yields a distinct estimate of the asymptotic covariance matrix. Griffiths, Hill, and Pope (1985) investigate the small sample properties of the two methods for the probit model and conclude that they differ negligibly.

3.3 Generalized Linear Models

3.3.1 Estimating GLIM

Nelder and Wedderburn (1972) and McCullough and Nelder (1983) explore a general class of statistical models which includes the classical normal regression model, log-linear model, probit and logit regression models, survival models and the Box-Cox (1964) transformed model to name just a few. These so-called generalized linear models (GLIM) share a common algorithm for the estimation of parameters by maximum likelihood which uses iterative weighted least squares with an adjusted dependent variable and resembles the Gauss-Newton in equation 3.2.13.

(a) Components of GLIM

Generalized linear models have three components.

1. A random component: The observable random variable y is assumed to be independently distributed with density function

$$f_y(y, \theta, \alpha) = \exp[c(y, \alpha) + (y h(\theta) - k(\theta)) / a(\alpha)]. \quad (3.3.1)$$

subject to $\partial k / \partial \theta_t = \theta_t \partial h / \partial \theta_t$ for all t . Thus, if α is known, then (3.3.1) is an exponential family of

distributions with canonical parameter θ . If α is unknown, then (3.3.1) may or may not be an exponential family.

2. The systematic component: The independent variables $x'_t = (x_{t1}, x_{t2}, \dots, x_{tK})$ produce what Nelder and Wedderburn call a linear predictor η given by

$$\eta_t = x'_t \beta \quad t=1, 2, \dots, T \quad (3.3.2)$$

where x_t is the t^{th} observation on K explanatory variables and β is a $K \times 1$ vector of unknown parameters.

3. A link function: The link function $g(\cdot)$ relates the random component to the systematic component of the model. In particular, the link function relates the linear predictor η_t to the expected value θ_t of the random variable y_t , or

$$\eta_t = g(\theta_t). \quad (3.3.3)$$

In terms of the density (3.3.1) note that $q(x'_t \beta) = g^{-1}$. For example, in the classical normal regression model the mean θ_t and the linear predictor η_t are equal to one another (i.e., $g(\theta_t) = \theta_t = E(y_t)$, $\eta_t = x'_t \beta$, and therefore $E(y_t) = x'_t \beta = \theta_t$).

However, if the dependent variable is binary, then the link function one which maps the unit interval $U(0,1)$ onto the real line. Call such a transformation

$$g(\pi_t) = \eta_t = x'_t \beta \quad (3.3.4)$$

where π is the (Bernoulli) probability of a success for an individual from the given population with characteristics x_i . Given this characterization, consider three frequently used transformations $g(\pi)$:

(i) Logit function

$$g_1(\pi) = \ln[\pi/(1-\pi)] \quad (3.3.5)$$

(ii) Probit function

$$g_2(\pi) = \Phi^{-1}(\pi) \quad (3.3.6)$$

(iii) Complementary Log-Log function

$$g_3(\pi) = \ln[-\ln(1-\pi)] \quad (3.3.7)$$

where Φ^{-1} is the inverse of the standard normal c.d.f. and the logit function is the inverse of a similar symmetric cumulative density function [see equations (3.2.16) and (3.2.17) above].

For many years the logit function was used as an approximation to the probit function because it is simpler to work with. However, given the availability of powerful, low-cost computers the probit function is no longer difficult to evaluate and offers several theoretical advantages over the logit transformation. Thus, the GLIM approach to estimating the probit model will be discussed in detail following a general presentation of the main principles of GLIM estimation.

(b) Mean and Variance

Let y_t be independently distributed with density

$$f_y(y, \theta, \alpha) = \exp[c(y, \alpha) + (yh(\theta) - k(\theta))/a(\alpha)] \quad (3.3.1)$$

where $\partial k / \partial \theta_t = \theta_t \partial h / \partial \theta_t$ for all t and $\theta_t \equiv q(x_t' \beta)$. Notice that as defined, θ_t varies across individual observations because of differences in the explanatory variables x_t , not because of differences in the underlying parameters β .

The regularity of the density function plays an important role in the derivation of the mean and variance of the random variable y . Therefore before proceeding, consider the following definition of regularity and a related proposition.

Definition 3.3.1 [Dhrymes (1974), p. 115] A probability density $f(y, \theta)$ is said to be regular if:

- (i) The range of the random variable y is independent of the parameter vector θ and
- (ii) the density $f(., \theta)$ possesses derivatives of at least third order with respect to θ , and these derivatives are bounded functions of y .

Given definition 3.3.1, consider the following proposition.

Proposition 3.3.1 [Bickel and Docksum (1977), p. 127] If

$$f_y(y, \theta, \alpha) = \exp[c(y, \alpha) + (yh(\theta) - k(\theta))/a(\alpha)]$$

is an exponential family and $h(\theta)/a(\alpha)$ has nonvanishing and continuous derivatives (with respect to θ) on the parameter space Θ , then the density is regular.

Proposition 3.3.1 establishes the regularity of the density (3.3.1) for every case considered below and enables one to derive the mean and variance of y in the following way.

Let y have probability density

$$f_y(y, \theta, \alpha) = \exp[c(y, \alpha) + (yh(\theta) - k(\theta))/a(\alpha)].$$

As a proper density function,

$$\int f_y(.) dy = 1. \quad (3.3.8)$$

Regularity of f_y allows derivatives to be carried through the integral, therefore, differentiating (3.3.8) with respect to θ yields

$$\int [\partial f_y / \partial \theta] dy = 0$$

or,

$$\int [\partial \ln(f_y) / \partial \theta] f_y dy = 0$$

where,

$$\ln f_y = \ln f_y(y, \theta, \alpha) = c(y, \alpha) + (yh(\theta) - k(\theta)) / a(\alpha). \quad (3.3.9)$$

This is equivalent to

$$E[\partial \ln(f_y) / \partial \theta] = 0.$$

Applying this fact to the density (3.3.1) yields

$$E \partial / \partial \theta \{ [yh(\theta) - k(\theta)] / a(\alpha) - c(y, \alpha) \} = 0$$

or,

$$E \left[[y(\partial h / \partial \theta) - (\partial k / \partial \theta)] / a(\alpha) \right] = 0. \quad (3.3.10)$$

Solving (3.3.10) for the expected value of y yields

$$E(y) = \frac{(\partial k / \partial \theta)}{(\partial h / \partial \theta)} = \frac{\theta (\partial h / \partial \theta)}{(\partial h / \partial \theta)} = \theta. \quad (3.3.11)$$

The variance of y can also be derived using the regularity property of f_y . First, take the second partial derivatives of (3.3.9) with respect to the parameter vector θ . This yields

$$\partial^2 \ln f_y / \partial \theta \partial \theta = [y(\partial^2 h / \partial \theta^2) - (\partial^2 k / \partial \theta^2)] / a(\alpha). \quad (3.3.12)$$

Using the fact that for regular density functions [see Hogg and Craig (1978), pp. 373-374]

$$E[\partial^2 \ln f_y / \partial \theta^2] = -E[\partial \ln f_y / \partial \theta]^2$$

one can square the left-hand-side of (3.3.10) and set the

negative expectation of this result equal to the expectation of (3.3.11). That is,

$$\begin{aligned}
 -E(\partial f_y / \partial \theta)^2 &= -E\{[y(\partial h / \partial \theta) - (\partial k / \partial \theta)] / a(\alpha)\}^2 \\
 &= -[a(\alpha)]^{-2} \{E[y(\partial h / \partial \theta)]^2 - 2E[y(\partial h / \partial \theta)(\partial k / \partial \theta)] \\
 &\quad + (\partial k / \partial \theta)^2\} \\
 &= -[a(\alpha)]^{-2} \{E[y^2(\partial h / \partial \theta)^2] - (\partial k / \partial \theta)^2\}
 \end{aligned}
 \tag{3.3.13}$$

By assumption, $\partial k / \partial \theta = \theta \partial h / \partial \theta$, therefore equation (3.3.13) may be rewritten as

$$\begin{aligned}
 -E(\partial f_y / \partial \theta)^2 &= -[a(\alpha)]^{-2} (\partial h / \partial \theta)^2 E(y^2) - [\theta (\partial h / \partial \theta)]^2 \\
 &= -[a(\alpha)]^{-2} (\partial h / \partial \theta)^2 [E(y^2) - \theta^2] \\
 &= -[a(\alpha)]^{-2} (\partial h / \partial \theta)^2 \text{Var}(y)
 \end{aligned}
 \tag{3.3.14}$$

Also, if $\partial k / \partial \theta = \theta \partial h / \partial \theta$, then

$$\partial^2 k / \partial \theta^2 = \partial h / \partial \theta + \theta \partial^2 h / \partial \theta^2.
 \tag{3.3.15}$$

Substituting (3.3.15) into (3.3.12) and taking the expectation yields

$$\begin{aligned}
 E(\partial^2 f_y / \partial \theta^2) &= [\theta (\partial^2 h / \partial \theta^2) - \theta (\partial^2 h / \partial \theta^2) - (\partial h / \partial \theta)] / a(\alpha) \\
 &= -[a(\alpha)]^{-1} (\partial h / \partial \theta).
 \end{aligned}
 \tag{3.3.16}$$

Equating (3.3.14) and (3.3.16) and solving for $\text{Var}(y)$ yields

$$\begin{aligned}
 [a(\alpha)]^{-2} (\partial h / \partial \theta)^2 \text{Var}(y) &= [a(\alpha)]^{-1} (\partial h / \partial \theta) \\
 \text{Var}(y) &= a(\alpha) (\partial h / \partial \theta)^{-1}.
 \end{aligned}
 \tag{3.3.17}$$

Notice that the variance of y depends on the canonical parameter θ and hence, on the mean of y . Also note that the variance of y depends on the dispersion parameter α (which, by assumption, is independent of θ).

The principle feature of the generalized linear model

is that for any density of the class (3.3.1), where the mean of y is a function of the linear predictor η , the maximum likelihood estimates of the parameters β can be obtained using iterative weighted least squares. This fact will be made clear in the following section where a general algorithm for estimating the parameters of a GLIM will be discussed.

(c) Algorithm for Fitting GLIM

An economist is seldom interested in $E(y)$ and $\text{Var}(y)$ alone. Instead, interest lies in estimating the effects of explanatory variables X_t on the mean of y_t (which is denoted as θ_t). In terms of the generalized linear model, it is the parameters of the linear predictor (β) which are of interest and an algorithm for finding their maximum likelihood estimates is desired.

The log-likelihood function associated with a random sample of size T from a density of the class (3.3.1) is denoted

$$\begin{aligned} \ell &= \ln \prod_{t=1}^T f(y_t, \theta_t, \alpha) = \sum_{t=1}^T \ln f(y_t, \theta_t, \alpha) \\ &= \sum_{t=1}^T \{ [y_t h(\theta_t) - k(\theta_t)] / a(\alpha) - c(y_t, \alpha) \} \end{aligned} \quad (3.3.18)$$

and is to be maximized with respect to β . Recalling that $\eta_t = x_t' \beta$, then for β_j ($j=1, 2, \dots, K$)

$$\partial \ell / \partial \beta_j = \sum [y_t \frac{\partial h}{\partial \eta} \frac{\partial \eta}{\partial \beta_j} - \frac{\partial k}{\partial \eta} \frac{\partial \eta}{\partial \beta_j}] a(\alpha)^{-1} \quad (3.3.19)$$

Also, recall the following assumptions:

1) $g^{-1} = q = q(\eta)$ and $\eta = X\beta$. This implies,

$$dq = (\partial q / \partial \eta) d\eta \text{ and therefore } dq/d\eta = \partial q / \partial \eta$$

2) $\partial k / \partial q = \theta \partial h / \partial q$.

Using assumptions 1 and 2, equation (3.3.19) may be written as

$$\begin{aligned} \partial \ell / \partial \beta &= \sum [y_t \frac{\partial h}{\partial q} \frac{dq}{d\eta} x_{tj} - \theta_t \frac{\partial h}{\partial q} \frac{dq}{d\eta} x_{tj}] a(\alpha)^{-1} \\ \partial \ell / \partial \beta_j &= \sum_1^T (y_t - \theta_t) (\partial h / \partial q) (dq/d\eta) x_{tj} / a(\alpha) \end{aligned} \quad (3.3.20)$$

and since $\text{Var}(y_t) = a(\alpha) / (\partial h / \partial q)$, then

$$\partial \ell / \partial \beta_j = \sum_1^T (y_t - \theta_t) (dq/d\eta) x_{tj} / \text{Var}(y_t). \quad (3.3.21)$$

Now, let

$$w_t \equiv [\text{Var}(y_t)]^{-1} (dq/d\eta)^2 \quad (3.3.22)$$

and the likelihood equations (3.3.21) become

$$\partial \ell / \partial \beta_j = \sum_1^T w_t (y_t - \theta_t) (d\eta / \partial q) x_{tj} = 0 \quad j=1, 2, \dots, K. \quad (3.3.23)$$

In order to use the Newton-Raphson or similar numerical optimization technique, the matrix of second derivatives (Hessian) is required. The jk^{th} element of the Hessian is

$$\begin{aligned} \partial^2 \ell / \partial \beta_j \partial \beta_k &= \sum_1^T (y_t - \theta_t) \partial / \partial \beta_k \{ [\text{Var}(y_t)]^{-1} (dq/d\eta) x_{tj} \} \\ &\quad - \sum_1^T \text{Var}(y_t)^{-1} (dq/d\eta)^2 x_{tj} x_{tk} \end{aligned} \quad (3.3.24)$$

Taking the expectation of (3.3.24) and recalling the fact

that $E(y_t) = \theta_t$, the jk^{th} element of the Information matrix becomes

$$\begin{aligned} I(\theta)_{jk} &= -E[H(\theta)_{jk}] = \sum_1^T [\text{Var}(y_t)]^{-1} (dq/d\eta)^2 x_{tj} x_{tk} \\ &= \sum_1^T W_t x_{tj} x_{tk} \end{aligned}$$

The Information matrix is formed by taking

$$I(\theta) = \sum_1^T W_t x_t x_t' = X'WX \quad (3.3.25)$$

where X is the $T \times K$ matrix of independent variables and $W = \text{diag}[W_1, W_2, \dots, W_T]$. The gradient vector may be formed by taking $g(\beta) = [\partial \ell / \partial \beta_1, \partial \ell / \partial \beta_2, \dots, \partial \ell / \partial \beta_K]'$ which is equal to

$$g(\beta) = \begin{bmatrix} \sum W_t (y_t - \theta_t) (d\eta / d\eta) x_{t1} \\ \sum W_t (y_t - \theta_t) (d\eta / d\eta) x_{t2} \\ \vdots \\ \sum W_t (y_t - \theta_t) (d\eta / d\eta) x_{tK} \end{bmatrix}$$

or,

$$g(\beta) = \sum_1^T W_t x_t (y_t - \theta_t) (d\eta / d\eta)_t \quad (3.3.26)$$

Substituting (3.3.25) and (3.3.26) into the method of scoring algorithm (3.2.2) yields

$$\begin{aligned} b_{n+1} &= b_n + I^{-1}(b_n) g(b_n) \\ b_{n+1} &= b_n + \left[\sum_1^T W_t x_t x_t' \right]^{-1} \sum_1^T W_t x_t (y_t - \theta_t) (d\eta / d\eta)_t \\ b_{n+1} &= \left[\sum_1^T W_t x_t x_t' \right]^{-1} \sum_1^T W_t x_t [X' b_n + (y_t - \theta_t) (d\eta / d\eta)_t] \end{aligned} \quad (3.3.27)$$

For the generalized linear model where $X_t' b_n = \eta_t$, (3.3.27) can be rewritten as

$$b_{n+1} = \left[\sum_1^T W_t x_t x_t' \right]^{-1} \sum_1^T W_t x_t [\eta_t + (y_t - \theta_t) (d\eta / d\eta_t)_t]. \quad (3.3.28)$$

Finally, if

$$\bar{y} \equiv \begin{cases} \eta_1 + (y_1 - \theta_1) (d\eta / d\eta_t)_1 \\ \eta_2 + (y_2 - \theta_2) (d\eta / d\eta_t)_2 \\ \vdots \\ \eta_T + (y_T - \theta_T) (d\eta / d\eta_t)_T \end{cases},$$

then in matrix notation,

$$b_{n+1} = (X' W X)^{-1} X' W \bar{y}. \quad (3.3.29)$$

Finally, since W is a diagonal matrix, $W = W^{k/2} W^{k/2}$, where $W^{k/2} = \text{diag}[W_1^{k/2}, W_2^{k/2}, \dots, W_T^{k/2}]$. Let $X^* = W^{k/2} X$ and $\bar{y}^* = W^{k/2} \bar{y}$ and (3.3.29) can be written as

$$b_{n+1} = (X^{*'} X^*)^{-1} X^{*'} \bar{y}^*. \quad (3.3.30)$$

Equation (3.3.30) resembles the FGLS estimator of equation (2.3.7). The major difference between the FGLS estimator (2.3.7) and the GLIM estimator (3.3.30) is that in the latter the regressors X^* and the dependent variable \bar{y}^* are functions of the n^{th} round estimates, b_n . Therefore, the algorithm for estimating the parameters of a generalized linear model is equivalent to using iterated feasible generalized least squares or, equivalently, iterated weighted least squares.

Now that the algorithm for estimating the parameters of a generalized linear model has been developed for the

most general case, consider the GLIM for a specific density of the class (3.3.1), namely, the binary choice model with the probit link function (3.3.6).

3.3.2 Example: Probit Regression Model

Let $\{y_1, y_2, \dots, y_T\}$ be a random sample of T Bernoulli trials with parameters π_t . The probability density function is denoted

$$f(y_t | \pi_t) = \pi_t^{y_t} (1 - \pi_t)^{1 - y_t} \quad t=1, 2, \dots, T.$$

Suppressing the t subscripts, the log of $f(y_t | \pi_t)$ is denoted

$$\ln f = y \ln(\pi) + (1 - y) \ln(1 - \pi) \quad (3.3.31)$$

collecting terms and rearranging yields

$$\ln f = y \ln[\pi / (1 - \pi)] + \ln(1 - \pi). \quad (3.3.32)$$

Finally, taking the inverse transformation of the natural logarithm yields

$$f(y | \pi) = \exp\{y \ln[\pi / (1 - \pi)] + \ln(1 - \pi)\}. \quad (3.3.33)$$

Proposition 3.3.2 The density (3.3.33) is of the class (3.3.1).

Proof: Recall the density specified by equation (3.3.1):

$$f_y(y, \theta, \alpha) = \exp[c(y, \alpha) + (y h(\theta) - k(\theta)) / a(\alpha)]. \quad (3.3.1)$$

Let

$$\begin{aligned} y &= y, & k(\theta) &= \ln[\pi / (1 - \pi)], \\ a(\alpha) &= 1, & c(y, \alpha) &= 0, \\ \theta &= \pi, \text{ and} & h(\theta) &= \ln[\pi / (1 - \pi)]. \end{aligned}$$

Substitution yields

$$f(y|\pi) = \exp\{y \ln[\pi/(1-\pi)] + \ln(1-\pi)\}$$

which is equivalent to (3.3.33). Now, consider the assumption $\partial k/\partial \theta = \partial \theta h/\partial \theta$. This implies

$$\partial k/\partial \theta = 1/(1-\pi) \text{ and}$$

$$\partial \theta h/\partial \theta = \pi [(1-\pi)/\pi] [(1-\pi+\pi)/(1-\pi)^2] = 1/(1-\pi).$$

■

A more general case of density (3.3.32) should be mentioned. In many instances, researchers are able to group observations. Thus, instead of observing a single Bernoulli trial, the researcher observes a sequence of n Bernoulli trials for each set of characteristics x_t . This gives rise to the binomial p.d.f.

$$f(y^*|\pi) = \binom{n}{y^*} \pi^{y^*} (1-\pi)^{n-y^*} \quad y^*=1,2,\dots,n$$

where $y^* = \sum_{i=1}^n y_i$. In terms of the density (3.3.1) this can be

expressed as

$$f(y^*|\pi) = \exp\{y^* \ln[\pi/(1-\pi)] + n \ln(1-\pi) + \ln\left(\binom{n}{y^*}\right)\}.$$

In this instance $c(y, \alpha) = \ln\left(\binom{n}{y^*}\right)$, $a(\alpha) = 1/n$, and $y = y^*/n$. Although this case will not be considered below, bear in mind that the method could just as easily be applied to situations where data can be grouped.

Returning to the example, suppose that (definition 3.2.1)

$$\pi_t = \Pr(y_t=1) = F(x_t'\beta) \quad t=1,2,\dots,T$$

and

$$\pi_t = F(x_t'\beta) = \Phi(x_t'\beta) \quad t=1,2,\dots,T.$$

The likelihood function is denoted

$$L = \prod_1^T \pi_t (x'_t \beta)^{y_t} [1 - \pi_t (x'_t \beta)]^{1-y_t} \quad (3.3.34)$$

and the log-likelihood is

$$\ell = \ln L = \sum_1^T y_t \ln(\pi_t) + (1-y_t) \ln(1-\pi_t) \quad (3.3.35)$$

or,

$$\ell = \sum_1^T y_t \ln[\pi_t / (1-\pi_t)] + \ln(1-\pi_t) \quad (3.3.36)$$

where $\pi_t = \pi(x'_t \beta)$. Using Proposition 3.3.1, the log-likelihood function (3.3.36) can be expressed in terms of the class of densities (3.3.1) where

$$\begin{aligned} \theta_t &= q(x'_t \beta) = \pi(x'_t \beta) \\ h(\theta_t) &= \ln[\pi_t / (1-\pi_t)], \\ k(\theta_t) &= \ln(1-\pi_t), \text{ and} \\ a(\alpha) &= 1. \end{aligned}$$

Recall from (3.3.17) that,

$$\text{Var}(y_t) = a(\alpha) / (\partial h / \partial q) = \pi_t (1-\pi_t). \quad (3.3.37)$$

In addition, note that

$$dq/d\eta = \phi(x'_t \beta) = (2\pi)^{-\frac{1}{2}} \exp\{-\frac{1}{2}(x'_t \beta)^2\}. \quad (3.3.38)$$

Substituting (3.3.37) and (3.3.38) into the weight function for the t^{th} observation (3.3.22) yields

$$w_t = \phi^2(x'_t \beta) / [\pi_t / (1-\pi_t)]. \quad (3.3.39)$$

The weight matrix may be formed by constructing a $T \times T$ matrix $W = \text{diag}[W_1, W_2, \dots, W_T]$ and the vector $\bar{y} = \eta + (y - \theta)(d\eta/dq)$ is formed by

$$\bar{y} = \pi^{-1} + (y - \pi)/\alpha$$

where $\pi_t = \pi(x'_t \beta)$, $\phi_t = d\pi_t/d\eta_t$ $t=1, 2, \dots, T$,

$$\Phi^{-1} = [\Phi^{-1}(x_1'\beta), \Phi^{-1}(x_2'\beta), \dots, \Phi^{-1}(x_T'\beta)]',$$

$$\Phi = [\Phi(x_1'\beta), \Phi(x_2'\beta), \dots, \Phi(x_T'\beta)]',$$

$$y = (y_1, y_2, \dots, y_T)', \text{ and}$$

$$(y - \Phi)/\phi = [(y_1 - \Phi_1)/\phi_1, (y_2 - \Phi_2)/\phi_2, \dots, (y_T - \Phi_T)/\phi_T]'. \quad (3.3.39)$$

From (3.3.29) the iterative equations for estimating the parameter vector β are

$$b_{n+1} = [X'W(b_n)X]^{-1}X'W(b_n)\bar{y}(b_n) \quad (3.3.40)$$

or,

$$b_{n+1} = [X^*{}'X^*]^{-1}X^*{}'\bar{y}^* \quad (3.3.41)$$

where $X^* = W^{\frac{k}{2}}X$, $\bar{y}^* = W^{\frac{k}{2}}\bar{y}$, and $W = \text{diag}[W_1^{\frac{k}{2}}, W_2^{\frac{k}{2}}, \dots, W_T^{\frac{k}{2}}]$. Iteration may be stopped once $\Phi_{t,n+1} - \Phi_{t,n}$ becomes sufficiently small for all $t=1, 2, \dots, T$.

Furthermore, it can be shown that

$$T^{\frac{k}{2}}(b - \beta) \sim N(0, T(X'WX)^{-1} + O_p(T^{-\frac{k}{2}})) \quad (3.3.42)$$

where $O_p(T^{\frac{k}{2}})$ means that the p.d.f. of $T^{\frac{k}{2}}(b - \beta)$ differs from the multivariate normal p.d.f. f_y by a term of order $T^{-\frac{k}{2}}$. The estimated covariance matrix, $(X'\hat{W}X)^{-1}$, is the inverse of the information matrix evaluated at the last round of estimates. It should be noted that although the method of scoring and the Gauss-Newton algorithms yield the same parameter estimates, the estimated covariance matrix is that yielded by the method of scoring; generally, it will not be equal to that obtained using the Gauss-Newton or Newton-Raphson.

In summary, the iterated feasible generalized least squares solutions to the maximum likelihood equations (3.3.41) are identical to those obtained by maximizing the

likelihood function using the Newton-Raphson technique. But, GLIM estimation requires use of first derivatives only. Caution is advised in interpreting the resulting estimate of the asymptotic covariance matrix, however. Although GLIM estimation resembles the Gauss-Newton algorithm in form, it implicitly uses the method of scoring in deriving estimates of the asymptotic covariance matrix. For a small sample comparison of various estimators of the asymptotic covariance matrix of the probit model, see Griffiths, Hill, and Pope (1987).

Even though the probit regression model is nonlinear in the parameters, its parameter estimates can be thought of as the result of iterated feasible generalized least squares (IFGLS). Given this interpretation, an algorithm for constructing a Stein-like shrinkage estimator for the probit model which uses the IFGLS probit estimates in place of the usual OLS estimates is suggested. In principle, the least squares interpretation is not important; the basic idea of Stein estimation is to shrink unrestricted maximum likelihood estimates toward a set of hypothesized values based on the value of the statistic used to test the restrictions. Unfortunately, the analytical risk properties of the resulting Stein-like estimator are as yet unknown. The empirical risk properties of a shrinkage estimator for the probit regression model will be examined in Chapter 6.

3.4 Computer Intensive Research Techniques

Efron (1979, p. 479) has said that the purpose of scientific theory is to reduce complicated situations to simpler ones. Nowhere is this principle more evident than in statistical theory. After all, the main goal of statistics is to separate out irrelevant information in data that obscures our understanding of the phenomenon under study. A classic example is embodied in the idea of a sufficient statistic for an unknown parameter θ ; once the value of the sufficient statistic is known, the data can tell us nothing more about the value of θ and therefore may be discarded. Simplification, it may be argued, is the hallmark of statistics.

Simplification in statistical theory can take many forms and these tend to change as research technology improves. Many techniques (the CORC procedure of Chapter 1, for instance) were developed and used because of their computational simplicity. With the arrival of powerful low-cost computers, many of these methods have been abandoned and replaced by ones more computationally demanding. The computer has led statisticians and others to redefine simplicity in the mathematical sciences. Two computer intensive tools available to econometricians are the Monte Carlo experiment and the bootstrap [Efron (1979)]. Both of these techniques can help shed light on problems which have proven to be analytically difficult. It is expected that their role in econometric research will

expand as practitioners seek information on the expanding body of promising yet little understood estimators of which the Stein-rule is but one member.

In the next two subsections, various aspects of Monte Carlo and bootstrapping will be defined and explored. The Monte Carlo technique will be used extensively in Chapter 5 to explore the size and coverage probabilities of confidence intervals and ellipsoids centered at the James-Stein estimator. In Chapter 6 it is again used to explore the risk properties of a Stein-like estimator of the binary choice regression model. The bootstrap is used in Chapter 5 to derive nonparametric confidence intervals and ellipsoids centered at Stein-rule parameter estimates.

3.4.1 Monte Carlo

Monte Carlo methods comprise a branch of applied mathematics concerned with experiments on random numbers. These methods have found the most widespread use in operations research and nuclear physics, but are also used in other fields of science like chemistry, biology, medicine, and statistics.

Monte Carlo methods can be applied to either deterministic or probabilistic mathematical problems depending on whether or not they are directly concerned with the behavior and outcome of random processes. In the simplest probabilistic Monte Carlo study, the researcher observes the behavior of random numbers generated in a way that simulates the random process of interest; the desired

information can be inferred from the behavior of the random numbers. In econometrics, random numbers are used to generate random samples from a population with a known probability density. These samples represent artificial data sets which are used to study the behavior of certain statistics of interest. For instance, Monte Carlo methods have been used to determine how various methods of computing the asymptotic covariance matrix of the probit model compare in finite samples [Griffiths, Hill, and Pope (1987)].

According to Hendry (1983), Monte Carlo experiments require one to define the data generation process, determine the feasible parameter space, define the relationship of interest, define the econometric technique to be investigated, and specify the object of study. In terms of an example, consider the simple regression model (2.1.1). The data are assumed to be generated by

$$y = X\beta + e \quad (3.4.1)$$

$$e \sim N(0, \sigma^2 I_T). \quad (3.4.2)$$

The total decision space of the experiment is $B \times \mathcal{T}$, where $B = \{\beta, \sigma^2 \mid \beta, \sigma \in R^1\}$ and $T \in \mathcal{T} = [T_0, T_1]$. Assume that knowledge is sought about an estimator of the the unknown parameters β ; therefore (3.4.1) may also be said to define (in this instance) the relationship of interest. In particular, suppose information is sought about the performance of the OLS estimator of β ; given this, the econometric estimator is $b = (X'X)^{-1}X'y$. Suppose that the object of the study is

to find the mean of the least squares estimator $E_e(b)$ where E_e denotes the expectation of the econometric estimator and where E_m will denote the expectation of the Monte Carlo estimator. Of course, this particular example is uninteresting because it is known that $E_e(b) = \beta$. A Monte Carlo study would be used only if $E_e(b)$ were difficult to evaluate analytically (or perhaps if E_e was a function of sample size or an unknown parameter).

Now suppose that $E_e(b)$ arises with some potential distribution of outcomes and the goal is to evaluate the econometric technique's ability to precisely locate the true mean β . $E_e(b)$ can be estimated by Monte Carlo simulation by generating a large number N of random samples (according to the data generating process (3.4.1) and (3.4.2)) and calculating a value of b for each sample. Taking the average of these using $N^{-1}\sum b_i = E_m(b)$ yields the Monte Carlo estimator of β , which is then compared to the true parameter value.

If $E_m(b)$ is considered for only a few points in the parameter space (i.e., for a few points in \mathbf{B} and for several values of T) then the study is referred to as a pilot study. Pilot studies are often useful for identifying peculiar regions of the parameter space which require more systematic investigation. On the other hand, if the claimed results hold over the entire parameter space, then the experiment is said to be valid. A reliable experiment is one which can be accurately reproduced using

a different set of random numbers from the same generation process.

Finally, it should be noted that the random numbers used in most Monte Carlo research are not actually random; they could be reproduced exactly if one knew the algorithm from which they were generated and the seed from which they were initialized. As a result, these numbers are called pseudo-random. All that is required is that they be statistically indistinguishable from actual random numbers generated by the assumed data generation process. Strictly speaking, even this characterization is not entirely accurate; nevertheless, for practical reasons one proceeds as if it were. [Hammersley and Handscoot (1964), p. 25]

3.4.2 The Bootstrap

Like the Monte Carlo, the bootstrap is a computer intensive research technique (technically speaking, it is the use of Monte Carlo to produce many bootstrap samples which makes it computer intensive). The most widely discussed approach to bootstrapping uses Monte Carlo simulation to estimate so-called bootstrap statistics based on the empirical distribution function of the random variable. The approximate c.d.f. is obtained by replacing the unknown c.d.f. with the empirical counterpart and then resampling the data to obtain a Monte Carlo distribution for the resulting random variable [Bickel and Freedman, (1981)].

The bootstrap is used most often to construct

confidence intervals when no other method is available. The process of constructing bootstrap confidence intervals is relatively simple to do and consists of executing the following steps. First, assume that $X = \{X_1, X_2, \dots, X_T\}$ is a random sample of size T from an unknown probability density F . The X_i may be residuals from a fitted model or from some other sequence of random variables. Now suppose that one wishes to make probability statements about some statistic $T(X)$. The statistic $T(X)$ depends on the sample size T , the functional form $T(X)$, and on the distribution F of the random variables X . Since F is unknown, it may be estimated by the empirical c.d.f. \hat{F} which, according to Freedman (1981), should be centered at the mean $\mu = T^{-1} \sum X_i$ and is found by assigning a mass of $1/T$ to each observation in X .

Next, draw a bootstrap sample $X^* = X^*_1, X^*_2, \dots, X^*_T$ by independently drawing with replacement T observations from \hat{F} . Thus, each new point is an independent random selection of the original T points. Notice that any individual element of the original data set may appear once, twice, etc., or not at all. At this point, the statistic of interest $T(X^*)$ is computed using the bootstrap sample. Then, a large number, N , of bootstrap samples are taken and $T(X^*)$ is computed for each one. One can now obtain an approximate c.d.f. of $T(X)$ by choosing some point $0 < c < N$ and calculating

$$\hat{F}(c) = \#\{T_i(X^*) \leq c\} / N \quad i=1, 2, \dots, T \quad (3.4.3)$$

which amounts to counting the number (#) of times $T(X^*)$ is less than c and dividing by the total number of samples taken.

Bickel and Freedman (1981) develop some asymptotic theory for the bootstrap, and, in a companion article, Freedman (1981) develops the theory as applied to regression models. In brief, Freedman shows that for the model 2.1.1 where variance is finite and $T^{-1}X'X \rightarrow C$, a finite matrix, the bootstrap approximation $T^{\frac{k}{2}}(\beta_b - b)$ closely approximates that of the usual asymptotic statistic $T^{\frac{k}{2}}(b - \beta)$, where β_b is the bootstrap estimate, b is the usual (MLE, OLS, method of moments, etc.) estimate, and β is the true parameter vector.

The approach is nonparametric since no distributional assumption is required in order to derive confidence intervals for any statistic. The procedure is intended for use in complex estimation problems where the sampling distribution of $T(X)$ depends on unknown values (or whose form may be unknown). However, caution is advised since Schenker (1985) finds that the bootstrap procedure fares quite poorly in a relatively "simple" situation. The method continues to draw controversy and should perhaps only be used when no reasonable alternatives are available.

This chapter has briefly summarized important results in time series estimation, nonlinear models, GLIM estimation, and computer intensive research techniques which are used in Chapters 4, 5, and 6.

Chapter 4
Improved Forecasts of Nominal GNP Growth
Using the St. Louis Equation

- 4.1 Introduction
- 4.2 The St. Louis Equation
- 4.3 The Statistical Model and Estimators
 - 4.3.1 Least Squares
 - 4.3.2 A General Family of Minimax Estimators
- 4.4 Data Analysis
- 4.5 Least Squares Estimation
- 4.6 Members of the General Family of Minimax Estimators
 - 4.6.1 Shrinking Toward the Origin
 - 4.6.2 Shrinking Toward the Sample Mean
 - 4.6.3 Shrinking Toward Hypotheses Implied by Model Selection
 - (a) Criterion Functions
 - (b) Estimates
 - 4.6.4 Shrinking Toward Economic Hypotheses
 - (a) Ahmed-Johannes Specification
 - (b) The Monetarist Hypothesis
 - (c) The Fiscalist Hypothesis
 - 4.6.5 Shrinking Towards the Principal Components Estimator
 - (a) Sequential Hypothesis Testing
 - (b) Rules-of-Thumb
 - (c) Model Selection
- 4.7 ARIMA Forecasts
 - 4.7.1 Forecasts and Forecast Error
 - 4.7.2 Hardware and Software
- 4.8 Results
 - 4.8.1 RMSE, 1962:2-1979:3, RLS vs. OLS Forecasts
 - 4.8.2 RMSE, 1962:2-1979:3, Shrinkage Forecasts
 - 4.8.3 RMSE, 1962:2-1982:3, RLS vs. OLS Forecasts
 - 4.8.4 RMSE, 1962:2-1982:3, Shrinkage Forecasts
- 4.9 Summary and Conclusion
 - A.4.1 Collinearity
 - A.4.2 Similarity of In-Sample and Out-of-Sample Regressor Matrices
 - A.4.3 Regression Diagnostics
 - A.4.3.1 Nonnormality
 - A.4.3.2 Autocorrelation
 - A.4.3.3 Heteroscedasticity

Chapter 4
Improved Forecasts of Nominal GNP Growth
Using the St. Louis Equation

4.1 Introduction

Economists and other empirical researchers often use linear regression equations to forecast the values of certain variables of interest. In practice, it is not known whether the use of uncertain prior information in linear regression models enhances or impairs forecasting performance. Therefore, a Stein-like estimator which shrinks unrestricted least squares estimates towards restricted least squares estimates based on the value of the statistic used to test the restrictions may provide a useful compromise between the two extreme cases. This proposition is explored below using a well-known reduced form macroeconomic model, the St. Louis equation.

Copas (1983) has investigated the forecasting problem using the classical normal linear regression model and found a shrinkage estimator similar to the one proposed by James and Stein (1961) which dominates the usual least squares predictor under mean square error of prediction loss. Copas's argument depends on several restrictive assumptions.

A1.1 The regression model is constant over past and future time periods.

A1.2 The residual error terms are independent and identically distributed normal random variables.

A1.3 The future regressor variables are 'similar' to

the past variables in the sense that they have the same mean and covariance structure.

Analytically, Jones and Copas (1986) explore the risk of the Copas (1983) predictor as A1.3 is relaxed and find it to be robust to small differences between appropriate moment matrices. Based on Monte Carlo evidence, they claim that the improved predictor can also be applied to situations where large and systematic differences between past and future samples are expected to exist. Unfortunately, they fail to compare the shrinkage predictor to any technique other than least squares.

Hill and Fomby (1986) also explore the out-of-sample forecast performance of several minimax and non-minimax shrinkage estimators using the mean square error of prediction norm and under various assumptions regarding the relationship between in-sample and out-of-sample regressors. Unlike Copas (1983) and Jones and Copas (1986), Hill and Fomby assume that the out-of-sample values of the regressors are nonstochastic. In effect, they replace A1.3 with a similar condition which pertains to the use of nonstochastic regressor variables.

On the basis of Monte Carlo evidence, Hill and Fomby conclude that for a certain general family of minimax predictors, the risk gain relative to the OLS estimator tends to be small since the minimax estimator converges to OLS rather quickly as the degree of collinearity increases. On the other hand, the potential risk improvement of non-

minimax rules from the same family appears to be substantial.

In this chapter, the forecast performance of several non-minimax shrinkage estimators will be studied using the well-known St. Louis equation [Andersen and Jordan (1968), Batten and Thornton (1983), (1984)]. At issue is whether uncertain prior information in the form of general linear hypothesis restrictions can be used to improve the forecast performance of a linear regression equation in realistic situations. Conditional root-mean-square forecast errors (RMSE's) are computed using actual and predicted values of the dependent variable obtained from OLS, RLS, and shrinkage regression equations.

The plan of the Chapter is as follows. In section 4.2 the St. Louis equation is introduced and in section 4.3 the economic model is recast into a statistical representation. In section 4.4, the data are discussed. The OLS estimates are presented, various members of the general family of minimax estimators are proposed, and estimates obtained in section 4.5. In section 4.6, a brief digression is made to discuss estimation of a univariate ARIMA forecasting model; the RMSE's of these forecasts are compared to those of OLS, RLS, and shrinkage forecasts for 1 to 16 step ahead forecast horizons in section 4.7. In the final section, a summary is given and conclusions are drawn.

4.2 The St. Louis Equation

Few single equation macroeconomic models have been studied more thoroughly than the Andersen-Jordan (1968) equation, which is otherwise known as the St. Louis equation. In its more recent forms [Batten and Thornton, (1983) and (1984)], the St. Louis equation is specified in the following way:

$$\dot{Y}_t = \alpha + \sum_{i=0}^k m_i \dot{M}_{t-i} + \sum_{i=0}^l g_i \dot{G}_{t-i} + e_t \quad (4.2.1)$$

where

\dot{Y}_t = quarterly observations of annualized growth rate of nominal GNP,

\dot{M}_t = annualized growth rate of a monetary aggregate (either M1 or monetary base),

\dot{G}_t = annualized growth rate of a federal fiscal aggregate (either high-employment federal expenditures or federal purchases of goods and services),

k, l = unknown parameters representing appropriate lag lengths for the monetary and fiscal variables,

m_i, g_i = unknown lag weights, and

e_t represents the cumulative effects of all other factors influencing the rate of GNP growth and is assumed to have zero mean.

Since (4.2.1) is to be used as a forecasting equation it is presumed that the growth rate of monetary and fiscal policy have an effect on the rate of growth of nominal GNP. Therefore, future values of the policy instruments \dot{M} and \dot{G}

should be useful in forecasting future rates of nominal GNP growth.

The unknowns l and k are assumed to be finite. Therefore, let $n = \max(l, k) < \infty$ be defined as the maximum lag length. In effect, this implies that one can choose values of l and k beyond which the values of g_i and m_i are known to be zero. These values of l and k are imposed on (4.2.1) before estimation. If the value of l or k chosen is too small, then estimates of g_i and m_i will be biased; if the value of l or k is too large, then estimates of g_i and m_i are inefficient. If l is too large and k too small (or, if k is too large and l too small), estimates of g_i and m_i may be inefficient and biased.

Batten and Thornton (1983) consider maximum lag lengths of $n=12$ and $n=16$, and let $l=k=n$, while Batten and Thornton (1984) consider maximum lag lengths of $n=8$, $n=12$, and $n=16$, again letting $l=k=n$.

The practice of equating l and k to the maximum lag length n is common; in essence, the maximum lag is chosen to include all lag weights for the variable with the longest distributed lagged effect and the maximum lag length for the other variable is then set equal to it. This procedure reflects a degree of caution on the part of the model builder, but it is not without its disadvantages. Even though it may be believed that this procedure is "costless" in the sense that no more of the available sample is used up as extra lagged values of the secondary

variable are included, arbitrarily setting $l=k$ is likely to add to the conditioning problem and to lower the power of hypothesis tests used for variable selection. In general, the unnecessary inclusion of irrelevant variables should be avoided if possible.

Estimates of the lag lengths appear to be quite sensitive to the sample on which they are based. In a previous study, Batten and Thornton (1984) conclude that k may be greater than 9 and that l may be greater than 8 based on the use of the FPE model selection criterion. If unbiased estimates of the lag weights or unbiased forecasts are sought, then it is important not to exclude relevant lag weights. The prudent model builder would insist that k and l be at least 10 and 9, respectively. And, since the two variables considered appear to have similar maximum lag lengths, equating l and k may in fact be warranted in this instance. Therefore, we choose $n=12$ and $l=k=n$.

Given that the maximum lag length for monetary and fiscal policy variables (k and l) is chosen to be 12, several possibilities emerge. On one hand the forecasting equation can be based on the unconstrained estimate of the lag weights, i.e., estimate $(\alpha \ m_0 \ m_1 \ \dots \ m_{12} \ g_0 \ \dots \ g_{12})'$. Or, one could base the forecasting equation on a more parsimonious specification, i.e., one that uses uncertain prior information in the form of general linear hypotheses about the parameters g_i and m_i . As in the choice of l and k , if the prior information is correct, then more efficient

unbiased forecasts are generated. However, if the information is incorrect, then the forecasts, though more efficient, will in general be biased. The amount of bias induced as a result of imposing incorrect restrictions on the parameters of the model will increase monotonically as hypothesis error increases.

In the following section, a formal statistical model of (4.2.1) is stated and several estimators discussed. Among these is a family of shrinkage estimators proposed by Mittelhammer and Young (1981) and extended by Mittelhammer (1984) which combines restricted and unrestricted estimators of the lag weights in a way that assures lower risk under mean square error of prediction norm.

4.3 The Statistical Model and Estimators

Consider the following statistical restatement of the St. Louis equation (4.2.1),

$$y = X\beta + e \quad (4.3.1)$$

where $y = \dot{Y}$ a $T \times 1$ vector of observable random variables,

$X = [1 \ \dot{M}_t \ \dot{M}_{t-1} \ \dots \ \dot{M}_{t-12} \ \dot{G}_t \ \dot{G}_{t-1} \ \dots \ \dot{G}_{t-12}]$ a $T \times K$ nonstochastic matrix of rank $K \leq T$,

$\beta = (\alpha \ m_0 \ m_1 \ \dots \ m_{12} \ g_0 \ g_1 \ \dots \ g_{12})'$ is a $K \times 1$ vector of unknown parameters

$K = 27$, and

$e \sim N(0, \sigma^2 I)$ is a $T \times 1$ vector of random disturbances.

Since the object of this study is to assess the out-of-sample forecasting performance of various estimators of (4.3.1), a model for the post-sample period is required.

Therefore, define X_0 to be an $N \times K$ matrix of future regressor values. Using assumptions A1.1 and A1.2, the statistical model for future values of y , denoted y_0 , can be stated as

$$y_0 = X_0 \beta + e_0 \quad (4.3.2)$$

where $e_0 \sim N(0, \sigma^2 I_N)$ and $E(e'e_0) = 0$. Like Hill and Fomby (1986), we take X_0 to be nonstochastic.

4.3.1 Least Squares

The ordinary least squares (OLS) and maximum likelihood estimator (MLE) of the parameter vector β is

$$b = (X'X)^{-1}X'y \sim N(\beta, \sigma^2(X'X)^{-1}).$$

This estimator is also the minimum variance unbiased estimator (m.v.u.e.) of β . The estimator

$$\hat{\sigma}^2 = (y - Xb)'(y - Xb) / (T - K) = s / (T - K)$$

is the m.v.u.e. of σ^2 , $s/\sigma^2 \sim \chi^2_{T-K}$, and σ^2 is independent of b .

4.3.2 A General Family of Minimax Estimators

The shrinkage estimator used below is developed by Mittelhammer and Young (1981) and extended by Mittelhammer (1984).

Let $\omega = R\beta - r = 0$ define a set of $J \leq K$ linear hypotheses representing uncertain prior information to be used in the estimation of β . The matrix R is $J \times K$, nonstochastic, $\text{rank}(R) = J$, and r is a $J \times 1$ vector of constants. The restricted least squares estimator of β is

$$b_r = b - S^{-1}R'(RS^{-1}R')^{-1}(Rb - r) \quad (4.3.3)$$

and the conventional statistic used to test the null

hypothesis $\omega=0$ against all alternatives is

$$u = (Rb-r)'(RS^{-1}R')^{-1}(Rb-r)/J\sigma^2 \sim F_{J, T-K, \lambda} \quad (4.3.4)$$

where the noncentrality parameter $\lambda = \omega'(RS^{-1}R')^{-1}\omega/2\sigma^2$.

Let \tilde{b} be an arbitrary estimator of the unknown vector β and let W be a symmetric positive definite matrix.

Weighted squared error loss is defined to be

$$L(\beta, \tilde{b}, W) = (\tilde{b} - \beta)'W(\tilde{b} - \beta). \quad (4.3.5)$$

An important special case of (4.3.5) is mean square error of prediction loss, which sets $W = X'X$. Out-of-sample mean squared error of prediction loss can be defined by letting $W = X_0'X_0$ and hence, $L(\beta, \tilde{b}, X_0'X_0)$.

The risk of using \tilde{b} to estimate β under weighted quadratic loss is defined to be average loss, i.e.,

$$E_{\beta}[L(\beta, \tilde{b}, W)] = E[(\tilde{b} - \beta)'W(\tilde{b} - \beta)]. \quad (4.3.6)$$

Mittelhammer (1984) has proposed an estimator which dominates (has risk no greater than) the maximum likelihood estimator of β under weighted quadratic loss.

Mittelhammer's estimator is Stein-like, combining sample with nonsample information ($\gamma=0$) in the following way

$$\delta = [1-c/u](b-b_r) + b_r \quad (4.3.7)$$

where $c=a(T-K)/J$,

$$0 < a < [2/(T-K+2)]\{\eta_L^{-1}\text{tr}[(RS^{-1}R')^{-1}RS^{-1}WS^{-1}R']-2\}, \quad (4.3.8.a)$$

W is a positive definite weight matrix, and η_L is the largest characteristic root of the expression in square brackets. A necessary condition for the dominance of δ over b is $J \geq 3$. For $W=X'X=S$ (in-sample mean square error of

prediction loss)

$$0 \leq a \leq 2(K-2)/(T-K+2). \quad (4.3.8.b)$$

The condition equivalent to (4.3.8.b) under which a shrinkage predictor is known to dominate the least squares predictor (using the out-of-sample prediction norm) is obtained by substituting $W=X_0'X_0$ into (4.3.8.a). This yields

$$0 \leq a \leq [2/(T-K+2)] \{n_L^{-1} \text{tr}[(RS^{-1}R')^{-1}RS^{-1}X_0'X_0S^{-1}R'] - 2\}. \quad (4.3.8.c)$$

The forecasting performance of this estimator has been studied by Hill and Fomby (1986) and found to offer little if any risk improvement over the OLS predictor under moderate degrees of multicollinearity. It turns out that the minimaxity condition (4.3.8.c) will seldom be met in practice; for instance, if $N < K$, W will not be positive definite and (4.3.8.c) fails.

Although the family of predictors based on the use of $W=X'X$ is not minimax under the out-of-sample prediction norm, it may, as Hill and Fomby (1986) suggest, perform well compared to the OLS predictor, even when A1.3 does not hold. In the spirit of Copas (1983) and Jones and Copas (1986) it is conjectured that use of the predictors based on (4.3.7) can result in substantial risk improvements compared to the least squares predictor over significant regions of the parameter space if A1.1 and A1.2 are satisfied, if (4.3.8.b) is met, and if A1.3 (or its nonstochastic equivalent) is not compromised too severely.

Unfortunately, these assumptions are hardly ever met in practice. Still, forecasters use regression methods to derive predictions of various economic phenomena. In this chapter, we demonstrate that even when these assumptions are violated, Stein-rule forecasts are generally much better than OLS forecasts and, in many instances, better than RLS and pretest forecasts.

4.4 Data Analysis

In this section the data and their source are discussed. In Appendix 4.1 the reader will find a brief discussion of the collinearity problem and in Appendix 4.2, a summary of a few crude measures of differences between the in- and out-of-sample data scatters.

Previous studies of the St. Louis equation are based on data which have since been revised. In December 1985, the U.S. Commerce Department announced a major revision of the U.S. National Income and Product Accounts (NIPA). Such revisions are made about every five years and reflect changes in definitions, classifications, and statistical treatment of economic source data. These revisions are made so that the NIPA more accurately reflect changes in the structure of the economy.

Carlson (1986) concludes that the impact of the 1985 revision on the estimated relationship between \dot{M} and \dot{Y} is slight. However, Carlson draws this conclusion by comparing regression equations of 4th quarter money growth on 4th quarter GNP growth estimated with revised and

unrevised data. Although the estimated money growth coefficient is similar across equations, M1 growth appears to be more highly correlated with GNP growth under the revision. Carlson makes no mention of the possible effects of the revision on fiscal policy measures. Unfortunately, no study has carefully documented the effects of the 1985 NIPA revision on the relationship between monetary and fiscal policy and GNP.

The data collected consist of quarterly observations on nominal GNP and actual federal purchases of final goods and services (denoted GGFE by Citibase) over the period 1959:1 to 1986:4. Narrowly defined money (M1) is available monthly; we use the quarterly average of monthly M1 as the monetary variable.

The data were obtained from the set of Citibase 5 $\frac{1}{4}$ " floppy diskettes available in June 1987. The estimation periods for the base forecasting equations are 1962:2 to 1979:3 and 1962:2 to 1982:3. These periods are convenient because they coincide with previous studies of the St. Louis equation. In addition, the end of each period marks a change in the Federal Reserve Board's (FRB) operating procedure. Based on this fact it is reasonable to argue that the underlying structural parameters in the reduced form equation are likely to change and that assumption A1.1 (the constancy of β) is violated. If this is true, one would not expect the usual regression based forecast methods to perform as well as they otherwise would. Such

structural changes are relatively easy to check ex post and a modest attempt to do so is made in the next section.

Changes in the data generation process are not the only ones of interest. Even if β is constant across regimes, differences in the relationship among the nonstochastic policy variables will affect forecast performance. Fomby and Hill (1988) have noted that prediction risk will be affected by differences in the location and orientation in the regressor space of in- and out-of-sample data scatters. They have studied the nature of such differences for the least squares predictor and suggest ways to assess the changes in risk for marginal deviations from a given value of X_0 . It is expected that the change in FRB operating procedure affected the level of and collinearity pattern associated with the exogenous variables themselves. By truncating the in-sample periods at each regime change we hope to highlight the robustness of the estimators examined to violations of assumption A1.1 and A1.3. We will return to discuss some of the specifics about these issues in the next section.

There is some controversy as to the proper choice of fiscal variable. Ahmed and Johannes (1984), Modigliani and Ando (1976), and Batten and Hafer (1983) use actual federal purchases while, Batten and Thornton (1983, 1984), Andersen and Jordan (1968), Schmidt and Waud (1973), and Seaks and Allen (1984) use high-employment expenditures (recent studies use federal cyclically adjusted budget

expenditures, FCABE). The results presented below do not appear to be sensitive to the choice of fiscal variable and therefore only those for actual purchases (GGFE) are reported.

Finally, the transformation used to approximate annualized percentage growth rate in the variables is the first difference of the natural logarithm times a factor of 400.

4.5 Least Squares Estimation

In this section, the unrestricted least squares estimates b are presented. In subsequent sections, specific members of the family (4.3.7), which shrink OLS estimates towards hypothesis restricted estimates, are discussed.

The least squares estimates of (4.3.1) for the sample period 1962:2-1979:3 are given below in Table 4.1. Note the large number of statistically insignificant parameter estimates (i.e., those with t -values below 2 for the two-tailed test at the 5% level). Also note that the coefficient estimates for m_7 , m_8 , m_{10} , g_6 and g_8 are individually significant at the 5% level. This is evidence that exclusion of lags weights shorter than 10 quarters for \dot{M} and 8 quarters for \dot{G} may result in biased least squares parameter estimates.

Least squares estimates for the 1962:2-1982:3 sample period, also appearing in Table 4.1, are roughly similar. In this case, the individually significant parameter

Table 4.1
OLS Estimates for the St. Louis Equation

Coefficient	1962:2 - 1979:3		1962:2 - 1982:3	
	OLS	Ahmed-Johannes	OLS	Ahmed-Johannes
α	2.22 (1.46)	2.53 (1.83)	3.36 (2.24)	2.81 (2.01)
m_0	0.59 (2.28)	0.54 (2.73)	0.59 (4.08)	0.44 (3.38)
m_1	0.41 (1.16)	-0.01 (0.05)	0.42 (2.81)	0.36 (2.80)
m_2	0.30 (0.83)	0.76 (2.77)	0.30 (1.90)	0.37 (2.74)
m_3	-0.36 (0.97)	-0.23 (0.83)	-0.13 (0.86)	-0.18 (1.29)
m_4	0.26 (0.69)	0.00 (0.03)	0.84 (0.55)	0.06 (0.42)
m_5	0.12 (0.31)	-	-0.20 (1.32)	-
m_6	0.08 (0.22)	-	0.19 (1.23)	-
m_7	-0.68 (2.20)	-	-0.17 (1.11)	-
m_8	0.60 (2.01)	-	0.02 (0.13)	-
m_9	0.14 (0.47)	-	0.38 (2.02)	-
m_{10}	-0.59 (2.19)	-	-0.68 (2.85)	-
m_{11}	0.15 (0.55)	-	0.43 (1.71)	-
m_{12}	0.14 (0.63)	-	-0.21 (0.95)	-
g_0	0.15 (2.84)	0.12 (2.63)	0.10 (2.03)	0.08 (1.74)
g_1	-0.02 (0.48)	-0.03 (0.64)	-0.04 (0.81)	-0.05 (1.00)
g_2	-0.06 (1.28)	-0.02 (0.42)	-0.03 (0.56)	-0.00 (0.03)
g_3	-0.05 (1.01)	-0.04 (0.87)	-0.06 (1.07)	-0.05 (1.05)
g_4	0.05 (0.86)	0.02 (0.55)	0.02 (0.48)	-0.01 (0.20)
g_5	-0.05 (0.81)	-	-0.02 (0.46)	-
g_6	0.13 (2.25)	-	0.08 (1.56)	-

Table 4.1 continued

Coefficient	1962:2 - 1979:3		1962:2 - 1982:3	
	OLS	Ahmed-Johannes	OLS	Ahmed-Johannes
g_7	0.05 (0.76)	-	0.02 (0.47)	-
g_8	-0.13 (2.38)	-	-0.13 (2.33)	-
g_9	-0.02 (0.50)	-	-0.02 (0.32)	-
g_{10}	0.02 (0.42)	-	-0.01 (0.20)	-
g_{11}	-0.01 (0.23)	-	-0.04 (0.68)	-
g_{12}	-0.01 (0.22)	-	0.00 (0.02)	-
D.W.	2.23	-	2.14	-
R^2	0.59	-	0.54	-
\bar{R}^2	0.34	-	0.32	-
Overall F Stat	2.38		2.45	
F-Test of Zero Restrictions	-	1.18	-	1.36

estimates appear 1 period further back in time. That is, m_9 , m_{10} , and m_{11} appear to be individually significant (or nearly so) at the usual test levels (5% or 10%).

The conclusion to be drawn from Table 4.1 is that inclusion of 12 lags of \dot{G} or \dot{M} does not appear to be unwarranted since unbiased estimates of β are sought. The effects of fiscal policy are small in magnitude and measured with relatively large error. Little else can be said, other than there is some evidence that fiscal policy growth occurring 1 quarter in the past has some positive lagged impact on the current growth rate of GNP between and that fiscal policy growth occurring 8 quarters in the past has a significant negative effect on current GNP growth. Other significant effects may be occurring; however, overparameterization in (4.3.1) may be impairing our ability to detect these statistically.

Having estimated (4.3.1) for each of the two sample periods under consideration, assumption A1.2 should be checked. This assumption is an important one in the derivation of the shrinkage estimator (4.3.7). Recalling that a necessary condition for minimaxity of δ is $b \sim N(\beta, \sigma^2(X'X)^{-1})$; note that for the model (4.3.1) this condition is implied by $e \sim N(0, \sigma^2 I)$. In Appendix 4.3 we report the results of diagnostic tests which confirm to a large degree this assumption.

Finally, we make a modest attempt to test for structural change between in-sample and out-of-sample

forecast periods. This cannot be done in an actual forecasting situation since post-sample values of the dependent variable are unknown. However, we can see if such a change occurs ex post and use this information in our assessment of the relative forecast performance of the rules considered.

Given enough observations in the post-sample period, one can estimate two sets of regression coefficients for in-sample and out-of-sample models and then test their equivalence. In order to detect a change in the relationship between X and y after 1979:3 we would use the combined sample 1962:2-1983:3, estimate separate slope and intercept parameters for the sub-samples 1962:2-1979:3 and 1979:4-1983:3, and then test their equivalence. Unfortunately there are only 16 observations in the 1979:4-1983:3 sub-sample and unique coefficient estimates cannot be obtained for both sets of slope parameters. Consequently, we use a test proposed by Fisher (1970) to detect changes in structure when one sub-sample contains more slope parameters than observations. Let SSE_r denote the sum of squared residuals from least squares estimation using the combined sample, T_1 the number of observations in the first partition (in-sample), T_2 the number of observations in the second partition (forecast sample), and \hat{e}_1 the vector of LS residuals obtained from the in-sample regression equation. The test statistic is

$$u = [(SSE_r - \hat{e}_1' \hat{e}_1) / T_2] / [\hat{e}_1' \hat{e}_1 / (T_1 - K)]$$

which has an F distribution with T_2 and $T_1 - K$ degrees of freedom if the null hypothesis of no structural change is true.

The test statistic for structural shift after 1979:3 is 11.98 which is distributed $F_{16,43}$ under the null hypothesis. The 5% critical value is 1.89 and the hypothesis is rejected. The test statistic for structural shift after 1982:3 is 2.61 which is distributed $F_{16,55}$ under the null hypothesis. The 5% critical value is 1.84 and the null hypothesis once again rejected. Thus, there is evidence to argue that $\beta \neq \beta_0$, arising as a consequence of the change in Federal Reserve Board intermediate monetary targets in October 1979 and 1982.

The robustness issue is important in forecasting since it is seldom known when the traditional relationships among the data have broken down until long after the break occurs. In terms of the forecasting performance of OLS based on (4.3.1) we know that the relationship between $X'X$ and $X_0'X_0$ (i.e., that between in-sample and out-of-sample regressors) changes somewhat for both of the estimation and forecast periods considered. Furthermore, it is likely that the relationship between β and β_0 also varies. Thus, the robustness of OLS and Stein-rule estimators is to be studied under two different types of misspecification.

4.6 Members of the General Family of Shrinkage Estimators

In this section several members of the general family of minimax estimators are discussed and estimated. These

include the James-Stein estimator (1961), the "Lindley estimator" (1961), estimators which make use of the sample and model selection criteria to obtain hypothesis restrictions, principal components estimators, and others.

4.6.1 Shrinking Toward the Origin

In the absence of any nonsample information it is sometimes suggested that shrinkage should be directed toward the origin; such an estimator shrinks the least squares estimates toward the hypothesis restriction that $\beta=0$ [see Chapter 2 above]. Thus, if $R=I_K$ and $r=0$ then the restricted least squares estimator (4.3.3) becomes $b_r=0$. The shrinkage estimator (4.3.7) reduces to the James-Stein (1961) estimator

$$\delta(JS) = [1 - as/b'Sb]b \quad (4.3.9)$$

where $s=(y-Xb)'(y-Xb)$, $S=X'X$, and $b_r(JS)=b_r=0$. The estimator $\delta(JS)$ dominates b under quadratic loss for $K \geq 3$ and for 'a' such that (4.3.8.a) is satisfied.

The estimator $\delta(JS)$ is seldom used because it is dominated by a rather simple modification, the positive-part rule $\delta(JS)^+$. A close examination of (4.3.9) indicates that if $as > b'Sb$, then $[1 - as/b'Sb] < 0$ and the algebraic sign of each element of $\delta(JS)$ is opposite that of b . In effect, the least squares estimates are being shrunk beyond their hypothesized values (zero), an unappealing event.

In response to this problem, the positive-part rule has been proposed. For the general family of shrinkage estimators (4.3.7), the positive-part rule sets $(1-c/u)=0$

when $c > u$. This of course is equivalent to setting $[1 - a_s/b'S_b] = 0$ when $a_s > b'S_b$ in (4.3.9). It can be shown that positive-part rules dominate the ordinary Stein-rules under quite general circumstances [Judge and Bock (1978)] and are used in most applications; explicit mention of the positive-part rules is henceforth omitted for expositional purposes since positive-part rules are the only ones used in this chapter.

4.6.2 Shrinking Toward the Sample Mean

Another possibility, inspired by Lindley (1962), is to shrink only the slope coefficients toward zero. We refer to this estimator as the "Lindley-rule" and denote it as $\delta(L)$; the Lindley-rule shrinks least squares estimates toward the hypothesis that $m_0 = m_1 = \dots = m_{12} = g_0 = \dots = g_{12} = 0$; the restricted least squares estimator, denoted $b_r(L)$, yields the sample mean, \bar{y} . In the absence of any other nonsample information about the past or future relationship between dependent and explanatory variables, a naive forecaster would typically choose the average value of the dependent variable to be the forecast value \hat{y}_0 , which is in this case average GNP growth. For this reason, it can be argued that for prediction the Lindley-rule is more appealing than the James-Stein rule (which also shrinks the intercept toward zero).

4.6.3 Shrinking Toward Hypotheses Implied by Model Selection

Another possibility is often pursued. Usually, the researcher is willing to admit that uncertainty exists about the number of lagged values of \dot{M} and \dot{G} to include as regressors in equation (4.3.1); starting from (4.3.1) with $l=k=12$, the researcher seeks a more parsimonious specification which is arrived at by using prespecified model selection rules to discriminate among competing models. The shrinkage estimator is obtained by shrinking least squares parameter estimates toward the hypothesis restrictions which emerge from the model selection procedure. In the following paragraphs several model selection rules are discussed which are thought to be particularly useful for specifying lag lengths [Geweke and Meese (1981)].

(a) Criterion Functions

As mentioned above, one approach to arriving at a more parsimonious specification of (4.3.1) is to choose $l < n$ and $k < n$ by minimizing a predetermined model selection criterion function [Geweke and Meese (1981)]. Model selection criterion functions trade goodness-of-fit for parsimony in specification of the model. The criterion contains a function of the sample variance, which decreases as regressors are added, and a penalty function, which increases as regressors are added. In order for the inclusion of an additional regressor to be considered an improvement in specification it must reduce the function of

sample variance by more than it increases the value of the penalty function.

Several such criterion functions are considered below. Akaike's information criterion (AIC) assumes the form

$$AIC(l,k) = \ln[(sse_{l,k})/T] + 2m/T \quad (4.6.1)$$

where $m = [1+2(l+1)+2(k+1)]$ is the number of regressors included, l is the number of lagged values of the fiscal variable to be included, k is the number of lagged monetary variables included, T is the number of observations, and $sse_{l,k}$ is the sum of squared errors from the regression equation based on the inclusion of l lagged fiscal and k lagged monetary variables. The criterion is to choose l and k such that

$$AIC(l,k) = \min\{AIC(l,k) \mid l,k = 0,1,\dots,n\} \quad (4.6.2)$$

where n is the value of the lag length beyond which it is certain that the lag weights are zero. Shibata (1981) has shown that AIC chooses the finite lag model that asymptotically minimizes sum of squared prediction errors.

If $l < n$ and $k < n$ are found to be the optimal lag lengths in the sense of (4.6.2), then the following set of hypothesis restrictions (to be applied to (4.3.1)) is implied:

$$m_{k+1} = \dots = m_{12} = g_{l+1} = \dots = g_{12} = 0.$$

The hypothesis restricted estimator is denoted $b_r(AIC)$ and the estimator of the family (4.3.7) which results from shrinking b towards $b_r(AIC)$ is denoted $\delta(AIC)$.

Another criterion which asymptotically minimizes sum

of squared prediction errors is Akaike's (1970) final prediction error (FPE) which takes the form

$$\text{FPE}(l,k) = [\text{sse}_{l,k}/T] (T + m)/(T - m). \quad (4.6.3)$$

The criterion is to choose l and k such that

$$\text{FPE}(l,k) = \min\{\text{FPE}(l,k) | l,k = 0,1,\dots,n\}. \quad (4.6.4)$$

Although FPE and AIC minimize the mean square error of prediction, it can be shown that each systematically over-predicts lag lengths and is therefore inconsistent. Two consistent criteria are the SBIC and the BEC criteria. The SBIC and BEC criteria are given below

$$\text{SBIC}(l,k) = \ln[(\text{sse}_{l,k})/T] + m \ln(T)/T \quad (4.6.5)$$

$$\text{BEC}(l,k) = (\text{sse}_{l,k})/T + m (\text{sse}_{n,n})/T \ln(T)/(T-K). \quad (4.6.6)$$

Once again, the object is to minimize these values over all values of l and k . The BEC and SBIC select lag lengths which are asymptotically neither too short nor too long. In small samples, however, Geweke and Meese (1983) provide some evidence that BIC and SBIC may yield specifications which contain fewer lagged variables than the AIC criteria. Hence, the fact that SBIC and BEC choose shorter lag lengths than FPE and AIC criteria for the St. Louis equation is hardly surprising.

As in the case of $b_r(\text{AIC})$ and the accompanying shrinkage estimator $\delta(\text{AIC})$, each of the models selected using $\text{FPE}(l,k)$, $\text{BEC}(l,k)$, and $\text{SBIC}(l,k)$ implies a set of hypothesis restrictions which is to be imposed on the unconstrained version of the St. Louis equation. Consequently, these rules give rise to the hypothesis

restricted estimators $b_r(\text{FPE})$, $b_r(\text{BEC})$, and $b_r(\text{SBIC})$ as well as an accompanying set of shrinkage estimators, denoted $\delta(\text{FPE})$, $\delta(\text{BEC})$, and $\delta(\text{SBIC})$.

(b) Estimates

Below, the results of the model selection process for the model (4.3.1) are reported. The use of the various model selection criteria yield similar models for the two sample periods under consideration. For the 62:2-82:3 sample period $b_r(\text{SBIC})=b_r(\text{FPE})=b_r(\text{BEC})=b_r(\text{AIC})$; each yields a specification which excludes all fiscal variables and contains contemporaneous and two lagged values of money growth. That is,

$$m_3=\dots=m_{12}=g_0=\dots=g_{12}=0.$$

For the 62:2-79:3 sample period minimization of SBIC, BEC, FPE, and AIC again yield an identical specification. In this instance, $l=1$ and $k=2$, meaning that contemporaneous and the first two lagged values of money growth are retained and contemporaneous and the first lagged value of fiscal growth are retained, i.e.,

$$m_3=\dots=m_{12}=g_0=\dots=g_{12}=0.$$

These results differ in some respects from those reported by Batten and Thornton (1983, 1984). Several reasons can be given. First, Batten and Thornton opt for a different measure of fiscal policy than the one used here; they use FCABE (Federal Cyclically Adjusted Budget Expenditures) rather than GGFE (actual purchases) as the measure of fiscal policy. Nevertheless, using BEC as a

model selection criterion, Batten and Thornton (1984) find $g_i=0$ for $i=0,1,\dots,12$ and $m_i=0$ for $i>1$ for the 62:2-82:3 sample period; using SBIC they find $g_i=0$ for $i\geq 0$ and $m_i=0$ for $i>2$. These results are identical to the ones obtained here.

A large difference is found when comparing models selected using FPE. Using the FPE criterion and unrevised data, Batten and Thornton (1984) select a model with $l=9$ and $k=10$. This is considerably different from the one selected in this study where the FPE criterion and revised data with GGFE as fiscal policy variable has been used.

To reconcile my results with those of Batten and Thornton (1984), models were selected using the revised measure of FCABE as the fiscal variable. The AIC and FPE criteria now yield a model where l and k are equal to 2. Using BEC and SBIC, fiscal policy is once again excluded and $k=2$. As a final check, models were selected using unrevised (pre-1985) data series like those available to Batten and Thornton. Using GGFE as the measure of fiscal policy, FPE and AIC select $k=12$ and $l=9$, a specification much like that of the previous study. Therefore, it seems likely that the differences in lag selection are due to the data revision rather than the choice of fiscal policy measure.

Examination of the revision, neatly summarized in the Survey of Current Business (December, 1985), indicates that the measure of federal government purchases was revised

upward. This was due mainly to a definitional change which adds the imputed value of a social insurances fund for military personnel to defense expenditures. The revisions appear to have had a marginally greater effect on the average annual rate of change of federal purchases for the 1972-1984 period than for GNP. Unfortunately, nothing is said in the Survey about quarterly fluctuations or about the effect of the revision on the growth rate of these aggregates over the 1959-1972 period. Nevertheless, in light of the evidence suggesting that the growth rate of fiscal policy have been revised upward to a greater extent than growth in GNP, one would expect the estimated slope parameters of the fiscal policy variable to be marginally closer to zero. Perhaps this accounts for the disappearance of long lags for fiscal policy under the new revision.

The model selection findings are summarized in Table 4.2 below, where F-statistics (for the test of the null hypothesis that the restrictions imposed are true against all alternatives) are reported along with the nominal p-values.

Table 4.2
Model Selection Estimators

Estimator $b_r(.)$	Sample	k	l	F- Stat	P- Value
Fiscal Variable = GGFE					
AIC, FPE, SBIC, BEC	62:2-79:3	2	0	1.00	.48
AIC, FPE, SBIC, BEC	62:2-82:3	2	-	1.19	.29
Fiscal Variable = FCABE					
AIC, FPE	62:2-79:3	2	2	.62	.87
SBIC, BEC	62:2-79:3	2	-	1.01	.47
Fiscal Variable = FCABE					
AIC, FPE	62:2-82:3	2	2	.95	.50
SBIC, BEC	62:2-82:3	2	-	1.27	.22
Unrevised Data: Fiscal Variable = GGFE					
AIC, FPE	62:2-82:3	12	9	.45	.80
SBIC, BEC	62:2-82:3	2	-	1.99*	.02
Batten and Thornton (1984) Fiscal Variable = FCABE					
FPE	62:2-82:3	10	9		
SBIC	62:2-82:3	2	-		
BEC	62:2-82:3	1	-		

*significant at the 5% level.

4.6.4 Shrinking Toward Economic Hypotheses

In many instances, economists are able to use economic theory as a source of hypothesis restrictions. When theory is available, its use can significantly improve the efficiency of estimation, but if inaccurate, the use of prior information may lead to a deterioration in risk performance. The family of shrinkage estimators (4.3.7) may be particularly useful in this respect since it enables an economist to combine theory with sample data in a way which results in lower risk under weighted quadratic loss than incurred if theory were ignored. Below, three sets of hypotheses are entertained which take advantage of this feature of Stein-rule estimation.

(a) The Ahmed-Johannes Hypothesis

Batten and Thornton (1983) estimate what they refer to as the "usual specification" of the St. Louis equation where \dot{Y} is regressed on contemporaneous and 4 lagged values of \dot{M} and \dot{G} . Thus, Batten and Thornton's usual specification implies the following restrictions on (4.3.1)

$$m_5 = \dots = m_{12} = g_5 = \dots = g_{12} = 0.$$

The "usual specification" is also the one used by Ahmed and Johannes (1984) and most other researchers. Thus, to avoid confusion, the restricted estimator will be denoted $b_r(AJ)$ and the shrinkage estimator $\delta(AJ)$. The restricted estimates $b_r(AJ)$ appear along with OLS estimates in Table 4.1.

Two subsidiary hypotheses are considered below which

use the Ahmed-Johannes specification as a starting point.

(b) The Monetarist Hypothesis

The first hypothesis considered, which shall be referred to as the Monetarist hypothesis (MH), assumes that the sum of the contemporaneous and first 4 lagged values of the monetary coefficients is equal to 1, that the sum of the contemporaneous and first 4 lagged values of the fiscal coefficients is equal to zero, and that all other lagged policy variables have no effect on GNP growth. Thus, the restrictions to be imposed on (4.3.1) are

$$m_5 = \dots = m_{12} = g_5 = \dots = g_{12} = 0 \quad (\text{AJ})$$

and

$$\sum_{i=0}^4 g_i = 0 \quad \text{and} \quad \sum_{i=0}^4 m_i = 1.$$

The hypothesis restricted estimator is denoted $b_r(\text{MH})$ and the resulting shrinkage estimator from the family (4.3.7) is $\delta(\text{MH})$.

(c) The Fiscalist Hypothesis

Another economist might insist that fiscal policy has a small positive effect on GNP growth and that the total effect of monetary policy on GNP growth is slightly overstated by the Monetarist Hypothesis. Call this the Fiscalist hypothesis (FH), which may be quantified by setting the sum of the contemporaneous and first 4 lagged values of the monetary coefficients is equal to 0.9 and the sum of the contemporaneous and first 4 lagged values of the fiscal coefficients is equal to 0.15, and all other lagged policy variables have no effect on GNP growth. The

resulting restrictions to be imposed on (4.3.1) by this hypothesis are

$$m_5 = \dots = m_{12} = g_5 = \dots = g_{12} = 0, \quad (AJ)$$

$$\sum_{i=0}^4 g_i = .15, \quad \text{and} \quad \sum_{i=0}^4 m_i = .9.$$

The hypothesis restricted estimator is denoted $b_r(FH)$. As before, the unrestricted least squares estimates of (4.3.1) are shrunk towards $b_r(FH)$ and the resulting Stein-like estimator is denoted as $\delta(FH)$.

4.6.5 Shrinking Towards the Principal Components Estimator

The principal components transformation is a means of establishing an orthogonal set of regressors which are ordered in a particularly useful way. The first principal component is the linear combination of the X_i , $i=1,2,\dots,K$, capturing the maximum amount of variation in the data in any one direction in the regressor space. The second component is another linear combination of the X_i and is chosen such that it captures the maximum amount of the remaining variation in the data, subject to the constraint that it be orthogonal to the first component and so on. When the data are nearly collinear, one or more of the principal components will capture little if any variation in the data (an indication of this is provided by the condition number of Belsley, Kuh, and Welsch (1980) which is a function of the ratio of largest root to each of the remaining roots). As a result, coefficient estimates for components with large condition numbers will be relatively

imprecise (or, in terms of t-tests for detecting differences from zero, they are often deemed to be insignificant). Therefore, it is thought that principal components which account for a small proportion of the variation in the data may be deleted without unduly reducing the explanatory power of the model.

More formally, let $P = (p_1 \ p_2 \ \dots \ p_K)$ denote the $K \times K$ matrix whose columns p_i are the orthonormal characteristic vectors of the regressor cross product matrix $X'X$ and let $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_K)$ be the diagonal matrix of corresponding characteristic roots. In addition, the characteristic roots have been ordered such that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_K$. That is to say, $X'XP = \Lambda P$ and $P'P = PP' = I_K$.

The principal components transformation of the linear model (4.3.1) is denoted

$$y = XPP'\beta + e = Z\theta + e \quad (4.6.7)$$

where $Z = (z_1 \ z_2 \ \dots \ z_K) = XP$ is the matrix of principal components and $\theta = P'\beta$ is the transformed parameter vector. In the transformed model (4.6.7), the regression variable $z_i = Xp_i$ is called the i^{th} principal component. The least squares estimator of θ in the reparameterized model is

$$\hat{\theta} = (Z'Z)^{-1}Z'y = \Lambda^{-1}Z'y \sim N(\theta, \sigma^2 \Lambda^{-1}). \quad (4.6.8)$$

Note also, that $\hat{\theta} = P'b$ and $P\hat{\theta} = b$.

Shrinkage estimators may be formed using (4.3.7) and (4.3.8.b) under the principal components reparameterization. According to the Mittelhammer (1984) result, the in-sample MSE of prediction risk associated

with use of the least squares principal components estimator (4.6.8) can be reduced by deleting at least 3 components. If the out-of-sample relationship among the regressor variables is similar to that of the in-sample regressors, then one can probably expect the forecasting equation based on the principal components regression to yield, on average, better forecasts than least squares (in the prediction mean square error sense). However, the precise conditions under which the use of $W = X'X \backslash X_0'X_0$ in (4.3.8.a) leads to improved forecasts are unknown. In addition, a procedure must be developed for determining the number of components to include in the estimation of the forecasting equation.

(a) Sequential Hypothesis Testing

One way to arrive at a more parsimonious specification of the reparameterized model is to delete irrelevant components based on a series of nested hypothesis tests. The principle is to test sequentially each of the following hypotheses

$$\begin{aligned} H_0: \theta_K &= 0 \\ H_0: \theta_{K-1} &= \theta_K = 0 \\ &\vdots \\ H_0: \theta_i &= \dots = \theta_{K-1} = \theta_K = 0 \end{aligned}$$

against all alternatives using some predetermined significance level. The sequence stops when a hypothesis is rejected. So, if the i^{th} hypothesis is rejected, then the model associated with the $(i-1)^{\text{th}}$ null hypothesis is

deemed to be the appropriate one.

Choosing an appropriate critical value for the pretest is problematic. As indicated in Chapter 2, the choice of significance level affects the risk properties of the pretest estimator [Judge and Bock (1978)]; the smaller the acceptable degree of type I error (smaller α), the greater the risk of the pretest estimator, other things equal. Optimal pretests have been suggested [Sawa and Hiromatsu (1973), Brook (1976)] by several authors. A minimax regret optimal critical value of 1.8 for a single hypothesis test is suggested by Sawa and Hiromatsu and for multiple hypotheses, Brook (1976) recommends minimax regret critical values ranging between 1.88 and 2.06. Brook (1976) presents a table of optimal critical values for orthogonal regressors (e.g., the principal components); these are employed below.

For the reparameterized model (4.6.7), the F-statistics for the test of the null hypothesis

$$H_0: \theta_{27} = 0$$

are 2.04 and 4.87 for the 1962:2-1979:3 and 1962:-1982:3 sample periods, respectively. Brook's optimal critical value is 1.88 and the first hypothesis is rejected for both samples. When no restrictions are imposed, the OLS, RLS, and shrinkage estimators will of course be equivalent.

(b) Rules-of-Thumb

Components are sometimes retained based on some pre-determined rule-of-thumb such as to retain those components

which account for at least 80% or 95% of the variation in the data.

Let $\theta(80)$ represent the hypothesis restricted estimator which arises by retaining the minimum number of principal components accounting for at least 80% of the variation in the data. The coefficients for the remaining components are restricted to equal zero.

Let $\theta(95)$ represent the hypothesis restricted estimator which arises by retaining the minimum number of principal components accounting for at least 95% of the variation in the data. The coefficients for the remaining components are restricted to equal zero.

Using these rules-of-thumb for the 1962:2-1979:3 sample, the first 8 principal components account for 81.8% of the variation in the data and the first 13 account for 95.5% of the variation in the data.

For $\theta(80)$ the restrictions on (4.6.7) are

$$\theta_9 = \theta_{10} = \dots = \theta_{27} = 0$$

and for $\theta(95)$ the restrictions to be imposed on (4.6.7) are

$$\theta_{14} = \theta_{15} = \dots = \theta_{27} = 0.$$

Using these rules-of-thumb for the 1962:2-1982:3 sample, the first 8 principal components account for 81.4% of the variation in the data and the first 14 account for 96.7% of the variation in the data.

For $\theta(80)$ the restrictions on (4.6.7) are

$$\theta_9 = \theta_{10} = \dots = \theta_{27} = 0$$

and for $\theta(95)$ the restrictions to be imposed on (4.6.7) are

$$\theta_{15} = \theta_{16} = \dots = \theta_{27} = 0.$$

A shrinkage estimator is formed using (4.3.7) and (4.3.8.b) under the principal components reparameterization. Again, the least squares estimator $\hat{\theta}$ is shrunk towards $\theta(80)$ and $\theta(95)$ based on the value of the statistic used to test the hypotheses restrictions. These rules are denoted $\delta(80)$ and $\delta(95)$, respectively.

(c) Model Selection

The model selection procedures discussed in section 4.6.4 may also be used to develop hypothesis restrictions for the principal components model. Starting from the fully parameterized model (4.6.7), let $\theta(AIC)$, $\theta(FPE)$, $\theta(BEC)$, and $\theta(SBIC)$ denote hypothesis restricted estimators arising from minimization of AIC, FPE, BEC, and SBIC, respectively.

In Table 4.3, the hypothesis restrictions implied by each criterion is summarized for the two samples considered.

Table 4.3
Model Selection Hypothesis Restrictions
for Principal Components

Estimator	Hypothesis Restrictions	
	62:2-79:3	62:2-82:3
$\theta(AIC)$	$\theta_{25} = \dots = \theta_{27} = 0$	$\theta_{27} = 0$
$\theta(FPE)$	$\theta_{24} = \dots = \theta_{27} = 0$	$\theta_{27} = 0$
$\theta(BEC)$	$\theta_{10} = \dots = \theta_{27} = 0$	$\theta_2 = \dots = \theta_{27} = 0$
$\theta(SBIC)$	$\theta_{10} = \dots = \theta_{27} = 0$	$\theta_{19} = \dots = \theta_{27} = 0$

It is interesting to note that although a consensus

model (i.e., one where all criteria choose the same specification) can be found under the original parameterization (4.3.1), no such model is obtained for the reparameterized model (4.6.7). Furthermore, because the range of possible models is quite wide, the number of components to delete in practice may be quite difficult to determine.

Note that the two Bayesian criteria select models with fewer parameters. For the 62:2-82:3 sample period SBIC retains only one component whereas BEC retains 18. FPE and AIC do not impose enough restrictions ($J > 2$) to make Stein-rule estimation useful; in-sample minimaxity cannot be assured and, like the sequential pretest estimator, these estimators are eliminated from consideration.

Stein-like estimators may be formed by shrinking $\hat{\theta} = P'b$ toward the model selection hypothesis restricted estimators $\theta(AIC)$, $\theta(FPE)$, $\theta(BEC)$, and $\theta(SBIC)$. Denote these as $\delta_{PC}(AIC)$, $\delta_{PC}(FPE)$, $\delta_{PC}(BEC)$, and $\delta_{PC}(SBIC)$, respectively.

Before discussing the issues surrounding forecast generation and measurement, a brief summary of the hypothesis restricted estimators is given in Table 4.4 and the symbols for all estimators considered below are presented in Table 4.5.

Table 4.4
Hypothesis Restricted Estimators

Estimator	Hypothesis Restrictions	
	62:2-79:3	62:2-82:3
$b_r(JS)$	$\alpha = m_0 = \dots = m_{12} = g_0 = \dots = g_{12} = 0$	same
$b_r(L)$	$m_0 = \dots = m_{12} = g_0 = \dots = g_{12} = 0$	same
$b_r(MS)^*$	$m_3 = \dots = m_{12} = g_1 = \dots = g_{12} = 0$	$m_3 = \dots = m_{12} = g_0 = \dots = g_{12} = 0$
$b_r(AJ)$	$m_5 = \dots = m_{12} = g_5 = \dots = g_{12} = 0$	same
$b_r(MH)$	$m_5 = \dots = m_{12} = g_5 = \dots = g_{12} = 0$ $\sum_{i=0}^4 m_i = 1, \sum_{i=0}^4 g_i = 0$	same
$b_r(FH)$	$m_5 = \dots = m_{12} = g_5 = \dots = g_{12} = 0$ $\sum_{i=0}^4 m_i = .9, \sum_{i=0}^4 g_i = .15$	same
$\theta(AIC)$	$\theta_{25} = \dots = \theta_{27} = 0$	not used
$\theta(FPE)$	$\theta_{24} = \dots = \theta_{27} = 0$	not used
$\theta(BEC)$	$\theta_{10} = \dots = \theta_{27} = 0$	$\theta_2 = \dots = \theta_{27} = 0$
$\theta(SBIC)$	$\theta_{10} = \dots = \theta_{27} = 0$	$\theta_{19} = \dots = \theta_{27} = 0$
$\theta(F)$	not used	not used
$\theta(80)$	$\theta_9 = \dots = \theta_{27} = 0$	$\theta_9 = \dots = \theta_{27} = 0$
$\theta(95)$	$\theta_{14} = \dots = \theta_{27} = 0$	$\theta_{15} = \dots = \theta_{27} = 0$

*Since all model selection criteria yield the same specification for model (3.1), the symbol $b_r(MS)$ is used to denote the restricted estimator implied by these criteria.

Table 4.5
Estimators and Symbols

Hypothesis	Restricted Estimator	Shrinkage Estimator
no restrictions	b	-
James-Stein	$b_r(JS)$	$\delta(JS)$
Lindley	$b_r(L)$	$\delta(L)$
[Mittelhammer Estimators]		
Ahmed-Johannes	$b_r(AJ)$	$\delta(AJ)$
Fiscal Hypothesis	$b_r(FH)$	$\delta(FH)$
Monetarist Hypothesis	$b_r(MH)$	$\delta(MH)$
Model Selection (MS) \neq FPE, BEC, AIC, SBIC	$b_r(MS)$	$\delta(MS)$
[Principal Components Estimators]		
Sequential F-Test	$\theta(F)$	$\delta(F)$
Model Selection (.) \neq FPE, BEC, AIC, SBIC	$\theta(.)$	$\delta_{PC}(.)$
Retain minimum which account for $\geq 80\%$ variation	$\theta(80)$	$\delta(80)$
Retain minimum which account for $\geq 95\%$ variation	$\theta(95)$	$\delta(95)$

4.7 ARIMA Forecasts

In order to get a better idea of how the various estimated models perform as forecasting equations, univariate ARIMA's were estimated for GNP using SAS's Proc ARIMA over both sample periods considered. Weak stationarity was achieved by taking the first difference of the natural logarithm of nominal GNP. Conveniently, this

transformation turns out to be the rate of growth transformation used in estimating (4.3.1).

For the 1962:2-1979:3 sample period, the following model was estimated using Ansley's (1979) maximum likelihood procedure:

$$\begin{matrix} (1-.2016L^4) & \dot{y}_t = .02158 + e_t \\ (.1212) & (.0014) \end{matrix}$$

where $\dot{y}_t = (1-L)\ln(y_t)$ and the estimated standard errors appear in parentheses.

For the 1962:2-1982:3 sample period, the estimated equation is:

$$\dot{y}_t = \begin{matrix} .0212 + e_t \\ (.0011) \end{matrix}$$

For this sample period the rate of growth of GNP appears to follow a random walk with a slight drift.

The Ljung-Box chi-square lack-of-fit test was used to check the residual series for departures from the null hypothesis of white noise. The statistic used is denoted

$$Q^* = T(T+2) \sum_{k=1}^M r_k^2 / (T-k) \sim \chi^2_{M-f}$$

where $r_k = \frac{\sum_{t=1+k}^M \hat{e}_t \hat{e}_{t-k}}{\sum_{t=1}^T \hat{e}_t^2}$ is the k^{th} autocorrelation

coefficient and f is the number of parameters (including the mean) estimated in the ARIMA model. The null hypothesis (e is white noise) cannot be rejected for either model. In Table 4.6 below, Q^* statistics are reported.

Table 4.6
ARIMA Diagnostics[†]

1962:2-1979:3			1962:2-1982:3		
To Lag	Q [*]	P-Value	To Lag	Q [*]	P-Value
6	4.56	.335	6	5.49	.359
12	8.80	.551	12	12.26	.345
18	13.37	.645	18	13.80	.681
24	16.04	.814	24	18.45	.733

[†] values generated by SAS Proc ARIMA.

4.7.1 Forecasts and Forecast Error

Given a vector of parameter estimates, call it \tilde{b} , the h step ahead forecast is derived by taking $\hat{y}_{0,h} = x_{T+h} \tilde{b}$, where x_{T+h} is the $1 \times K$ vector of future values of the explanatory variables in time $T+h$, and \hat{y}_0 is the $h \times 1$ vector of predicted nominal GNP growth.

Forecast error for the h step ahead forecast horizon is measured using root-mean-square error (RMSE), which is defined to be:

$$RMSE(h) = [h^{-1} \sum_{i=1}^h (y_{0,i} - \hat{y}_{0,i})^2]^{1/2}.$$

4.7.2 Hardware and Software

Computer programs for all estimators (except ARIMA) were written in SAS, Version 5.16, Matrix programming language and run on an IBM 3081 mainframe computer. The ARIMA models were estimated using SAS's Proc ARIMA.

4.8 Results

In this section, the root-mean-square errors (RMSE) for 1 to 16 step ahead out-of-sample forecasts generated from the equations estimated above are compared to those of the Stein-rule shrinkage estimator (4.3.7).

The comparison of RMSE's can be conducted on several levels. One could compare the 1 to 16 step ahead RMSE's of each RLS predictor to those of the unrestricted least squares predictor. In addition, one could see how each of the Stein predictors compares to the two predictors which it combines (i.e., OLS and RLS). Finally, one could see how the RMSE's of the Stein predictors compare to other predictors, i.e., ones not based on the linear regression model. To serve this last purpose, the Stein-rule forecasts are to be compared to univariate ARIMA forecasts.

4.8.1 RMSE, 1962:2-1979:3, RLS vs. OLS Forecasts (Table 4.7)

In Table 4.7 several of the RLS predictors discussed in section 4.7 are compared to the unrestricted least squares predictor b. The tabled values are obtained by dividing the RMSE yielded by each of the RLS predictors by that of the OLS predictor. Numerical values greater than 1 indicate that the average OLS forecast is lower in RMSE than the average RLS forecast over the h step ahead horizon. In the first column of Table 4.7 the RMSE of the ARIMA is divided by the RMSE of the OLS predictor.

The ARIMA forecast and those of $b_r(L)$ are clearly

superior to all other 1 step ahead forecasting equations. In fact, the RMSE of the 1 step ahead ARIMA forecast is about 1/10 that of the OLS forecast and the RMSE of $b_r(L)$ is 1/5 that of OLS. Of all the rules considered, only $\theta(AIC)$ (which is denoted $\theta(AC)$) is worse in terms of RMSE than OLS for the 1 step ahead forecast.

The OLS forecasting equation tends to perform better than the RLS equations for the 3-8 step ahead horizons, the exception being $\theta(AC)$. For the 9 step horizon and beyond, the OLS predictor performs poorly while $b_r(L)$ and the ARIMA perform quite well. In fact, the ARIMA and $b_r(L)$ have the same RMSE over the 16 step horizon for the 1979:4-1983:3 forecast period.

4.8.2 RMSE, 1979:4-1983:3, Shrinkage Forecasts (Tables 4.8-4.11)

To assess the forecast performance of the various forecast equations considered, the 1 to 16 period ahead RMSE's of least squares, restricted least squares, and ARIMA models are compared to a Stein-like shrinkage predictor for the equations estimated with the 1962:2-1979:3 sample.

Several of the shrinkage predictors discussed in section 4.7 are compared to the unrestricted least squares predictor b in Table 4.8. The tabled values are obtained by dividing the RMSE yielded by each shrinkage predictor by that of the OLS predictor. Numerical values greater than 1 indicate that the RMSE of the OLS forecast is lower than that of the shrinkage forecast for the h step ahead

horizon. In the first column, the RMSE associated with the OLS predictor is presented.

Table 4.8 is strikingly similar to Table 4.7. The reason for this can be seen by looking at Table 4.9 which compares the RMSE of each shrinkage predictor to its RLS component, again by taking the ratio of the Stein rule RMSE to that of its RLS counterpart. Using the positive part rule assures that when the value of the statistic used to test the hypothesis restrictions is sufficiently small (indicating 'good' hypotheses), the Stein and RLS parameter estimates and forecasts will be equivalent. This occurs for each of the shrinkage estimators considered except $\delta(\text{JS})$ and $\delta(\text{L})$. For these predictors, $\delta(\text{L})$ is a better forecaster than $b_r(\text{L})$ for the 2-12 step ahead forecast horizons but not for the 13-16 period ahead horizons. At the 16 step ahead horizon, the RMSE of $\delta(\text{L})$ is only 3% greater than that of $b_r(\text{L})$ and only 4% greater than that of the ARIMA.

In terms of RMSE, the Lindley rule forecasts are quite good. In Figure 4.1, the actual values of GNP growth are plotted along with forecasts from $b_r(\text{L})$, $\delta(\text{L})$, the ARIMA, and b . Notice how closely b is able to predict GNP growth up until the second quarter of 1981. Beginning with 81:3, the OLS forecasts become quite erratic (indicative of the unusual behavior of velocity about that time). Note also that the shrinkage forecasts are less variable than OLS forecasts. The Lindley-rule forecasts are being shrunk

toward the average value of GNP growth (the horizontal line) and are actually a linear combination of X_0b and \bar{y} . Once the least squares predictions go awry, the RMSE's of $\delta(L)$ fall below those of OLS (i.e., for steps 9 and beyond).

The problem with the mean forecasting rule $b_r(L)$ (and the ARIMA in this instance) should be apparent. With this estimator, one is unable to forecast turning points or directions of change in GNP growth. If the OLS equation correctly predicts the direction of change but is highly variable, then use of $\delta(L)$ can offer a significant advantage over other rules (like $b_r(L)$ and ARIMA) which ignore this important dimension of the forecasting problem. Using the Lindley-rule $\delta(L)$ will reduce variance and preserve the additional predictability available from models which try to use economic theory as a forecasting tool. Unfortunately, the OLS forecasts predict growth in the wrong direction for all but three of the last 8 forecast periods. The Lindley-rule mimics these directional changes because it is merely a linear combination of X_0b and \bar{y} . As such, it predicts the same directional change as least squares with lower variance.

4.8.3 RMSE, 1962:2-1982:3, RLS vs. OLS Forecasts (Table 4.11)

In Table 4.11, several of the RLS predictors discussed in section 4.7 are compared to the unrestricted LS predictor. Once again, the tabled values are obtained by

dividing the RMSE of each RLS forecast equation by that of the OLS equation.

For the 1962:2-1982:3 sample period, least squares is consistently out-performed by all other forecast equations in terms of RMSE. The best one step ahead forecast equation is again provided by the ARIMA, and is closely followed by $b_r(JS)$ and $b_r(L)$. It should be noted that the economic hypothesis forecast equations provided by $b_r(FH)$ and $b_r(MH)$ do poorly, as does the model selection rule(s) $b_r(MS)$.

For the 4 step ahead horizon the ARIMA is again the best forecaster, followed by $b_r(L)$ and $\theta(95)$. For the 8 step ahead horizon, $b_r(L)$ is on average the best forecaster, while the ARIMA is second best and $\theta(95)$ is the third best in terms of RMSE. Over the entire 16 period forecast period, the top performers are the ARIMA and $b_r(L)$, each of which has RMSE that is approximately 1/3 that of the OLS estimator.

4.8.4 RMSE, 1982:4-1986:3, Shrinkage Forecasts (Tables 4.12-4.15)

To assess the forecast performance of the shrinkage predictors, the least squares, restricted least squares, and ARIMA are compared to Stein-like shrinkage forecast equations which are estimated using the 1962:2-1982:3 sample period. In Table 4.12 several of the shrinkage predictors discussed in section 4.6 are compared to the unrestricted OLS predictor b . The tabled values are obtained by dividing the RMSE of each shrinkage predictor

by that of the OLS predictor. Once again, numerical values greater than 1 indicate that the OLS predictor out performs the RLS predictor. In the first column, the RMSE's of the OLS forecasts are presented.

Table 4.12 is similar to Table 4.11. The reason for this can be seen by looking at Table 4.13 which compares the RMSE of each shrinkage predictor to its RLS component. Again, this is accomplished by taking the ratio of the RMSE's produced by the Stein and the RLS predictors. Use of the positive part rule ensures that when the value of the statistic used to test the hypothesis restrictions is sufficiently small (indicating 'good' hypotheses) the Stein and RLS parameter estimates and forecasts will be equivalent. In Table 4.13 maximum shrinkage occurs under the positive part rule for $\delta(AJ)$, $\delta(MH)$, $\delta(FH)$, and $\theta_{PC}(BEC)$. Of the remaining rules, only $\delta(JS)$ has lower RMSE than its RLS component $b_r(JS)$.

The degree of improvement of b_r over δ does not appear to be large in most cases. In only 5 instances (all for the James-Stein predictor) does the RMSE of the shrinkage predictor not fall between that of the OLS and RLS predictors from which it is comprised. In these five cases, the RMSE of the shrinkage predictor is actually below that of the restricted and unrestricted predictors.

Although the ARIMA (see Tables 4.11 and 4.14) performs best overall, the Lindley rule $\delta(L)$ out performs every RLS predictor except $b_r(L)$. For the 16 step ahead horizon, the

RMSE of $\delta(L)$ is only 37% higher than that of the ARIMA.

Of all the models considered, those which do not attempt to use monetary and fiscal policy to forecast GNP growth ($b_T(L)$ and ARIMA) yield the best forecasts. Note that the RMSE's of $\delta(L)$ are well below those of the OLS predictor (2.3 to 2.7 times smaller); and, though the RMSE of $\delta(L)$ is twice that of $b_T(L)$ for the one step ahead forecast, it is only 1.2 times that of $b_T(L)$ at steps 9-11 and is on average 1.36 times that of $b_T(L)$ over the entire period.

In Figure 4.2, the actual values of GNP growth are plotted along with forecasts from $b_T(L)$, $\delta(L)$, the ARIMA, and b . Even though the OLS forecasts are volatile, they forecast changes in GNP growth in the correct direction more often than not (9 of 15 times). The Lindley rule $\delta(L)$ can be expected to perform well in this case. Forecast variance will be reduced and turn points will be foreseen. The 37% percent sacrifice in RMSE (compared to ARIMA) may be a small price to pay for better predictability of directional changes.

4.9 Summary and Conclusion

In summary, $\delta(L)$ appears to have many advantages as a forecasting equation of stationary time series. This rule essentially allows one to combine mean forecasts with those of an explanatory model. The ARIMA, which for the 1962:2-1982:3 period is a random walk with drift, forecasts a

slight upward trend. The enlightened economist would in most instances prefer to use economic theory to assist in the generation of GNP forecasts. Theory would be especially useful during periods of abnormal growth, like the 79:4-82:4 period. For instance, suppose that following a period of rapid money growth inflationary pressure rises and nominal GNP is expected to grow faster than the normal rate. The ARIMA cannot hope to use this information in the way that $\delta(L)$ can. Using $\delta(L)$ would enable us to shrink forecasts from an explanatory model toward the average value of the variable of interest. If past money growth has had no effect on GNP growth, then shrinking parameter estimates toward the mean would be greater, and the forecasts would reflect this. If on the other hand money growth really does have a strong effect, the forecasts are pulled away from the mean towards the values implied by the estimated econometric model.

Under the positive-part Stein-rule with maximum shrinkage, forecasts frequently converge to RLS forecasts when the restrictions are supported by the sample. In this chapter, the RLS forecasts were by and large superior to the OLS forecasts and consequently, so were the Stein-rule forecasts. In terms of the robustness of various prediction rules, of those which depend on assumptions about the constancy of the regression equation and similarity between in- and out-of-sample regressor matrices, the Stein-rule estimators perform much better

than least squares. The ARIMA, which does not depend on these assumptions, performs best overall, but is closely rivaled by the Lindley-rule.

Finding shrinkage rules which effectively forecast nonstationary (explanatory) time series models may be problematic because of the difficulty selecting an appropriate set of hypotheses towards which to shrink. None of the procedures investigated here can be deemed an unqualified success. Of all the hypothesis restriction schemes examined in this chapter, those based on model selection and economic theory performed poorly relative to naive forecast equations like ARIMA and mean forecasting rules.

As a final note, it is interesting that of the RLS rules based on economic theory, the Monetarist hypothesis appears to be the best. The additional restrictions that the sum of the first 5 money coefficients equal unity and the first 5 government purchases coefficients equal zero improved the predictability of the Ahmed-Johannes specification.

NOTES

¹Several other estimators were used for the 62:2-82:3 sample period, but are not reported. These include the Ridge [Hoerl and Kennard (1970a)], Iterative Ridge [Hoerl, Kennard and Baldwin (1975)], a truncated Stein-like estimator [Dey and Berger (1983)], polynomial distributed lag estimators, and a new variant of a principal components estimator.

(1) The iterative ridge estimator performed very well indeed compared to the Lindley rule.

(2) Two sets of PDL restrictions were chosen based on model selection. The following specifications were suggested.

	Current + 12 Lags M	Current + 12 Lags G	Method of Selection
Polynomial	3	2	BEC, SBIC
Degrees	6	2	AIC, FPE
	4	4	(Adkins Prior)

The last choice, polynomials of degree 4 for money and purchases growth, is my best guess at what the degrees might be. The performance of these methods leaves much to be desired. In fact, they performed much the same as the models arising from model selection reported in the text.

(3) Finally, a principal components-like estimator was created which performed very well, overall. In this estimator, separate sets of principal components were formed for monetary and fiscal policy variables after the data had been centered. The intercept was estimated as a

residual. The restrictions imposed were such that only one component of money and one of purchases were allowed to affect GNP growth. The effect then is to shrink toward the average value of the dependent variable while letting the two major components account for deviations from the mean. Given the success of the Lindley rule, the performance of this estimator is not surprising. Unfortunately, there is no precedent for using such an estimator as its analytical risk has not been studied. The problem here is that it is uncertain how shrinkage affects estimation of the intercept, as it is done here. Some additional research is required, but the problem appears to be both tractable and interesting.

Table 4.7

RMSE Comparison Between RLS and OLS Estimators
 Estimation Period: 1962:2 - 1979:3
 Forecast Period: 1979:4 - 1983:3

Forecast	ARIMA/ OLS	JS	L	MS	AJ	MH	FH	$\theta(AC)$	$\theta(BC)$	$\theta(95)$
1 step	0.10	0.91	0.19	0.53	0.56	0.50	0.60	1.00	0.85	0.78
2 step	0.58	1.82	0.52	0.58	0.65	0.56	0.70	1.06	0.87	0.81
3 step	1.18	1.82	1.23	0.78	0.67	0.57	0.77	1.06	1.92	1.86
4 step	1.19	2.16	1.23	0.97	0.85	0.74	0.90	1.08	1.92	1.86
5 step	1.60	3.00	1.58	1.57	1.70	1.72	1.63	1.07	1.93	1.88
6 step	2.03	3.75	1.96	1.57	1.72	1.73	1.67	1.12	2.04	2.00
7 step	1.61	2.98	1.57	1.43	1.45	1.43	1.45	0.88	1.79	1.81
8 step	1.45	2.80	1.40	1.28	1.35	1.36	1.36	0.79	1.61	1.61
9 step	0.95	1.74	0.94	0.93	1.00	0.97	1.02	1.30	1.23	1.22
10 step	0.96	1.57	0.96	0.99	1.02	0.99	1.06	1.19	1.20	1.21
11 step	0.72	1.20	0.72	0.74	0.77	0.75	0.79	1.03	0.90	0.91
12 step	0.62	1.00	0.63	0.70	0.73	0.71	0.76	1.00	0.84	0.84
13 step	0.49	0.78	0.50	0.58	0.61	0.59	0.63	1.03	0.69	0.69
14 step	0.49	0.79	0.49	0.59	0.61	0.59	0.63	1.02	0.69	0.69
15 step	0.50	0.84	0.50	0.60	0.63	0.60	0.64	1.02	0.69	0.69
16 step	0.49	0.85	0.49	0.60	0.62	0.59	0.63	1.02	0.68	0.68

Note: In the column labeled "ARIMA/OLS" the RMSE of the ARIMA estimator has been divided by that of the OLS estimator and is reported for each of the forecast horizons. In the remainder of the columns we report the ratios of the RMSE of the RLS estimator to that of OLS, i.e., $JS = RMSE[b_r(JS)] / RMSE[OLS]$. Also note $\theta(AC)$ represents $\theta(AIC)$ and $\theta(BC)$ represents both $\theta(SBIC)$ and $\theta(BEC)$.

Table 4.8

RMSE Comparison Between Stein-Rules and OLS Estimator

Estimation Period: 1962:2 - 1979:3

Forecast Period: 1979:4 - 1983:3

Forecast	RMSE OLS	JS	L	MS	AJ	MH	FH	$\delta(AC)$	$\delta(BC)$	$\delta(95)$
1 step	7.69	0.84	0.38	0.53	0.56	0.50	0.60	1.00	0.85	0.78
2 step	5.51	0.83	0.50	0.58	0.65	0.56	0.70	1.06	0.87	0.81
3 step	4.50	0.83	1.00	0.78	0.67	0.57	0.77	1.06	1.92	1.86
4 step	3.90	0.83	1.00	0.97	0.85	0.74	0.90	1.08	1.92	1.86
5 step	3.51	0.87	1.27	1.57	1.70	1.72	1.63	1.07	1.93	1.88
6 step	3.21	0.90	1.57	1.57	1.72	1.73	1.67	1.12	2.04	2.00
7 step	3.80	0.93	1.23	1.43	1.45	1.43	1.45	0.88	1.79	1.81
8 step	4.01	0.90	1.09	1.28	1.35	1.36	1.36	0.79	1.61	1.61
9 step	6.10	0.90	0.82	0.93	1.00	0.97	1.02	1.30	1.23	1.22
10 step	6.39	0.90	0.86	0.99	1.02	0.99	1.06	1.19	1.20	1.21
11 step	8.16	0.92	0.64	0.74	0.77	0.75	0.79	1.03	0.90	0.91
12 step	9.35	0.92	0.60	0.70	0.73	0.71	0.76	1.00	0.84	0.84
13 step	11.64	0.91	0.51	0.58	0.61	0.59	0.63	1.03	0.69	0.69
14 step	11.32	0.91	0.51	0.59	0.61	0.59	0.63	1.02	0.69	0.69
15 step	10.93	0.91	0.52	0.60	0.63	0.60	0.64	1.02	0.69	0.69
16 step	10.81	0.91	0.51	0.60	0.62	0.59	0.63	1.02	0.68	0.68

Note: In the column labeled "RMSE OLS," the RMSE of the OLS estimator is reported for each of the forecast horizons. In the remainder of the columns we report the ratios of the RMSE of the Stein estimator to that of the OLS estimator. i.e., $JS = RMSE[\delta(JS)] / RMSE[b]$. Also note $\delta(AC)$ represents $\delta(AIC)$ and $\delta(BC)$ represents both $\delta(SBIC)$ and $\delta(BEC)$.

Table 4.9

RMSE Comparison Between Stein-Rules and RLS Estimators
 Estimation Period: 1962:2 - 1979:3
 Forecast Period: 1979:4 - 1983:3

Forecast	JS	L	MS	AJ	MH	FH	$\delta(AC)$	$\delta(BC)$	$\delta(95)$
1 step	0.92	1.91	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2 step	0.45	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3 step	0.45	0.81	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4 step	0.38	0.81	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5 step	0.29	0.80	1.00	1.00	1.00	1.00	1.00	1.00	1.00
6 step	0.24	0.80	1.00	1.00	1.00	1.00	1.00	1.00	1.00
7 step	0.31	0.78	1.00	1.00	1.00	1.00	1.00	1.00	1.00
8 step	0.32	0.77	1.00	1.00	1.00	1.00	1.00	1.00	1.00
9 step	0.51	0.86	1.00	1.00	1.00	1.00	1.00	1.00	1.00
10 step	0.57	0.89	1.00	1.00	1.00	1.00	1.00	1.00	1.00
11 step	0.77	0.89	1.00	1.00	1.00	1.00	1.00	1.00	1.00
12 step	0.91	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00
13 step	1.16	1.03	1.00	1.00	1.00	1.00	1.00	1.00	1.00
14 step	1.15	1.03	1.00	1.00	1.00	1.00	1.00	1.00	1.00
15 step	1.08	1.02	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16 step	1.07	1.02	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Note: The numbers in the table are obtained by dividing the RMSE of the Stein Estimator by that of the restricted estimator, i.e., $JS = RMSE[b_r(JS)]/RMSE[\delta(JS)]$. Also note $\delta(AC)$ represents $\delta(AIC)$ and $\delta(BC)$ represents both $\delta(SBIC)$ and $\delta(BEC)$.

Table 4.10

RMSE Comparison Between Stein-Rules and the ARIMA Model
 Estimation Period: 1962:2 - 1979:3
 Forecast Period: 1979:4 - 1983:3

Forecast	OLS/ ARIMA	JS	L	MS	AJ	MH	FH	$\delta(AC)$	$\delta(BC)$	$\delta(95)$
1 step	9.66	8.17	3.69	5.13	5.48	4.91	5.80	9.68	8.23	7.54
2 step	1.70	1.42	0.86	0.99	1.11	0.97	1.19	1.81	1.49	1.38
3 step	0.84	0.70	0.85	0.66	0.56	0.48	0.65	0.89	1.62	1.57
4 step	0.83	0.70	0.84	0.81	0.71	0.62	0.76	0.90	1.61	1.56
5 step	0.62	0.54	0.79	0.98	1.06	1.07	1.01	0.67	1.20	1.17
6 step	0.49	0.44	0.77	0.77	0.84	0.85	0.82	0.55	1.00	0.98
7 step	0.61	0.58	0.76	0.88	0.89	0.89	0.90	0.54	1.10	1.12
8 step	0.68	0.62	0.75	0.88	0.93	0.93	0.94	0.54	1.11	1.11
9 step	1.04	0.94	0.85	0.97	1.04	1.02	1.07	1.35	1.28	1.27
10 step	1.04	0.94	0.89	1.03	1.06	1.03	1.10	1.24	1.25	1.26
11 step	1.38	1.28	0.89	1.02	1.06	1.03	1.10	1.42	1.25	1.26
12 step	1.59	1.46	0.96	1.11	1.17	1.13	1.21	1.59	1.34	1.34
13 step	2.02	1.85	1.05	1.19	1.23	1.19	1.28	2.09	1.41	1.40
14 step	2.03	1.86	1.05	1.21	1.25	1.21	1.29	2.09	1.41	1.41
15 step	1.99	1.83	1.03	1.21	1.25	1.21	1.29	2.05	1.39	1.38
16 step	2.03	1.86	1.04	1.23	1.26	1.21	1.29	2.09	1.39	1.40

Note: In the column labeled "OLS/ARIMA" the RMSE of the ARIMA estimator has been divided by that of the OLS estimator and is reported for each of the forecast horizons. In the remainder of the columns we report the ratios of the RMSE of the Stein estimator to that of the ARIMA, i.e., $JS = RMSE[\delta(JS)]/RMSE[ARIMA]$. Also note $\delta(AC)$ represents $\delta(AIC)$ and $\delta(BC)$ represents both $\delta(SBIC)$ and $\delta(BEC)$.

Table 4.11

RMSE Comparison Between RLS and OLS Estimators
 Estimation Period: 1962:2 - 1982:3
 Forecast Period: 1982:4 - 1986:3

Forecast		ARIMA/									
		OLS	JS	L	MS	AJ	MH	FH	θ (SB)	θ (BC)	θ (95)
1	step	0.16	0.17	0.18	0.36	0.37	0.38	0.42	0.40	0.42	0.28
2	step	0.17	0.32	0.19	0.52	0.49	0.49	0.51	0.45	0.45	0.27
3	step	0.23	0.58	0.24	0.52	0.49	0.48	0.50	0.43	0.45	0.26
4	step	0.23	0.66	0.23	0.55	0.48	0.48	0.49	0.43	0.51	0.31
5	step	0.25	0.77	0.25	0.53	0.46	0.46	0.47	0.42	0.49	0.29
6	step	0.33	0.93	0.32	0.54	0.47	0.46	0.50	0.46	0.49	0.29
7	step	0.30	0.90	0.30	0.50	0.45	0.45	0.48	0.43	0.52	0.37
8	step	0.30	0.89	0.30	0.49	0.44	0.44	0.47	0.44	0.50	0.36
9	step	0.32	0.90	0.32	0.50	0.46	0.45	0.49	0.46	0.50	0.37
10	step	0.33	0.92	0.33	0.51	0.46	0.45	0.49	0.47	0.50	0.40
11	step	0.34	0.94	0.34	0.55	0.50	0.50	0.55	0.48	0.54	0.48
12	step	0.31	0.87	0.31	0.56	0.55	0.55	0.59	0.46	0.60	0.50
13	step	0.29	0.80	0.29	0.58	0.55	0.55	0.60	0.44	0.71	0.56
14	step	0.29	0.81	0.30	0.62	0.57	0.57	0.60	0.45	0.74	0.57
15	step	0.32	0.78	0.32	0.68	0.64	0.64	0.66	0.47	0.82	0.65
16	step	0.31	0.76	0.31	0.71	0.65	0.64	0.66	0.46	0.86	0.68

Note: θ (SBIC) and θ (BEC) are abbreviated by θ (SB) and θ (BC), respectively. The columns are formed by taking the ratio of the RMSE of the Stein-Rule to that of the RLS estimator, i.e., $JS = RMSE[\delta(JS)] / RMSE[b_r(JS)]$.

Table 4.12

RMSE Comparison Between Stein-Rules and OLS Estimators

Estimation Period: 1962:2 - 1982:3

Forecast Period: 1982:4 - 1986:3

		RMSE									
Forecast		OLS	JS	L	MS	AJ	MH	FH	δ (SB)	δ (BC)	δ (95)
1	step	23.29	0.89	0.39	0.36	0.37	0.38	0.42	0.78	0.42	0.32
2	step	16.95	0.90	0.37	0.52	0.49	0.49	0.51	0.76	0.45	0.31
3	step	14.36	0.89	0.36	0.52	0.49	0.48	0.50	0.75	0.45	0.30
4	step	12.79	0.88	0.36	0.55	0.48	0.48	0.49	0.75	0.51	0.34
5	step	11.93	0.90	0.38	0.53	0.46	0.46	0.47	0.75	0.49	0.32
6	step	10.92	0.90	0.41	0.54	0.47	0.46	0.50	0.75	0.49	0.32
7	step	10.95	0.89	0.40	0.50	0.45	0.45	0.48	0.75	0.52	0.39
8	step	10.56	0.90	0.39	0.49	0.44	0.44	0.47	0.73	0.50	0.38
9	step	10.00	0.90	0.40	0.50	0.46	0.45	0.49	0.74	0.50	0.39
10	step	9.53	0.90	0.40	0.51	0.46	0.45	0.49	0.74	0.50	0.42
11	step	9.13	0.89	0.41	0.55	0.50	0.50	0.55	0.75	0.54	0.49
12	step	9.64	0.89	0.40	0.56	0.55	0.55	0.59	0.75	0.60	0.51
13	step	10.28	0.89	0.40	0.58	0.55	0.55	0.60	0.75	0.71	0.57
14	step	9.92	0.89	0.40	0.62	0.57	0.57	0.60	0.75	0.74	0.58
15	step	9.98	0.89	0.43	0.68	0.64	0.64	0.66	0.76	0.82	0.65
16	step	10.22	0.88	0.42	0.71	0.65	0.64	0.66	0.76	0.86	0.69

Note: δ (SBIC) and δ (BEC) are abbreviated by δ (SB) and δ (BC), respectively. The columns are formed by taking the ratio of the RMSE of the Stein-Rule to that of the RLS estimator, i.e., $JS = RMSE[\delta(JS)/RMSE[b_r(JS)]$. Column one is the RMSE of OLS.

Table 4.13

RMSE Comparison Between Stein-Rule and RLS Estimators
 Estimation Period: 1962:2 - 1982:3
 Forecast Period: 1982:4 - 1986:3

Forecast	JS	L	MS	AJ	MH	FH	$\delta(\text{SB})$	$\delta(\text{BC})$	$\delta(95)$
1 step	5.04	2.12	1.00	1.00	1.00	1.00	1.93	1.00	1.14
2 step	2.78	1.94	1.00	1.00	1.00	1.00	1.67	1.00	1.13
3 step	1.53	1.52	1.00	1.00	1.00	1.00	1.71	1.00	1.12
4 step	1.33	1.52	1.00	1.00	1.00	1.00	1.73	1.00	1.09
5 step	1.16	1.48	1.00	1.00	1.00	1.00	1.77	1.00	1.09
6 step	0.97	1.27	1.00	1.00	1.00	1.00	1.64	1.00	1.09
7 step	0.99	1.32	1.00	1.00	1.00	1.00	1.72	1.00	1.06
8 step	1.00	1.27	1.00	1.00	1.00	1.00	1.67	1.00	1.06
9 step	0.99	1.23	1.00	1.00	1.00	1.00	1.58	1.00	1.06
10 step	0.97	1.23	1.00	1.00	1.00	1.00	1.57	1.00	1.05
11 step	0.95	1.21	1.00	1.00	1.00	1.00	1.53	1.00	1.02
12 step	1.02	1.28	1.00	1.00	1.00	1.00	1.61	1.00	1.02
13 step	1.10	1.36	1.00	1.00	1.00	1.00	1.68	1.00	1.02
14 step	1.08	1.35	1.00	1.00	1.00	1.00	1.66	1.00	1.02
15 step	1.12	1.32	1.00	1.00	1.00	1.00	1.60	1.00	1.01
16 step	1.16	1.36	1.00	1.00	1.00	1.00	1.65	1.00	1.01

Note: $\delta(\text{SBIC})$ and $\delta(\text{BEC})$ are abbreviated by $\delta(\text{SB})$ and $\delta(\text{BC})$, respectively. The columns are formed by taking the ratio of the RMSE of the Stein-Rule to that of the RLS estimator, i.e., $\text{JS} = \text{RMSE}[\delta(\text{JS}) / \text{RMSE}[b_r(\text{JS})]$.

Table 4.14

RMSE Comparison Between Stein-Rules and ARIMA Estimators
 Estimation Period: 1962:2 - 1982:3
 Forecast Period: 1982:4 - 1986:4

Forecast	OLS/ ARIMA	JS	L	MS	AJ	MH	FH	$\delta(\text{SB})$	$\delta(\text{BC})$	$\delta(95)$
1 step	5.90	5.29	2.30	2.16	2.23	2.26	2.51	4.62	2.48	1.90
2 step	5.67	5.12	2.15	2.96	2.79	2.78	2.91	4.32	2.59	1.77
3 step	4.21	3.76	1.55	2.20	2.07	2.05	2.13	3.16	1.89	1.27
4 step	4.28	3.80	1.54	2.37	2.08	2.06	2.12	3.22	2.19	1.46
5 step	3.85	3.47	1.47	2.05	1.80	1.77	1.85	2.89	1.90	1.27
6 step	2.99	2.71	1.25	1.62	1.41	1.40	1.50	2.26	1.47	0.98
7 step	3.23	2.90	1.30	1.63	1.48	1.47	1.56	2.43	1.69	1.29
8 step	3.23	2.91	1.26	1.60	1.44	1.43	1.52	2.38	1.64	1.25
9 step	3.04	2.74	1.23	1.53	1.40	1.39	1.49	2.26	1.54	1.20
10 step	3.01	2.71	1.23	1.54	1.38	1.38	1.49	2.24	1.52	1.28
11 step	2.92	2.62	1.22	1.61	1.48	1.48	1.61	2.19	1.58	1.45
12 step	3.18	2.84	1.29	1.81	1.76	1.75	1.88	2.39	1.91	1.64
13 step	3.41	3.04	1.37	1.98	1.91	1.90	2.05	2.57	2.44	1.97
14 step	3.34	2.97	1.36	2.09	1.91	1.90	2.03	2.52	2.49	1.96
15 step	3.08	2.74	1.33	2.11	2.00	1.97	2.05	2.36	2.53	2.03
16 step	3.20	2.85	1.37	2.27	2.11	2.08	2.12	2.45	2.76	2.22

Note: $\delta(\text{SBIC})$ and $\delta(\text{BEC})$ are abbreviated by $\delta(\text{SB})$ and $\delta(\text{BE})$, respectively. The columns are formed by taking the ratio of the RMSE of the Stein-Rule to that of the RLS estimator, i.e., $\text{JS} = \text{RMSE}[\delta(\text{JS}) / \text{RMSE}[b_r(\text{JS})]$.

FORECASTED VS. ACTUAL VALUES OF GNP GROWTH

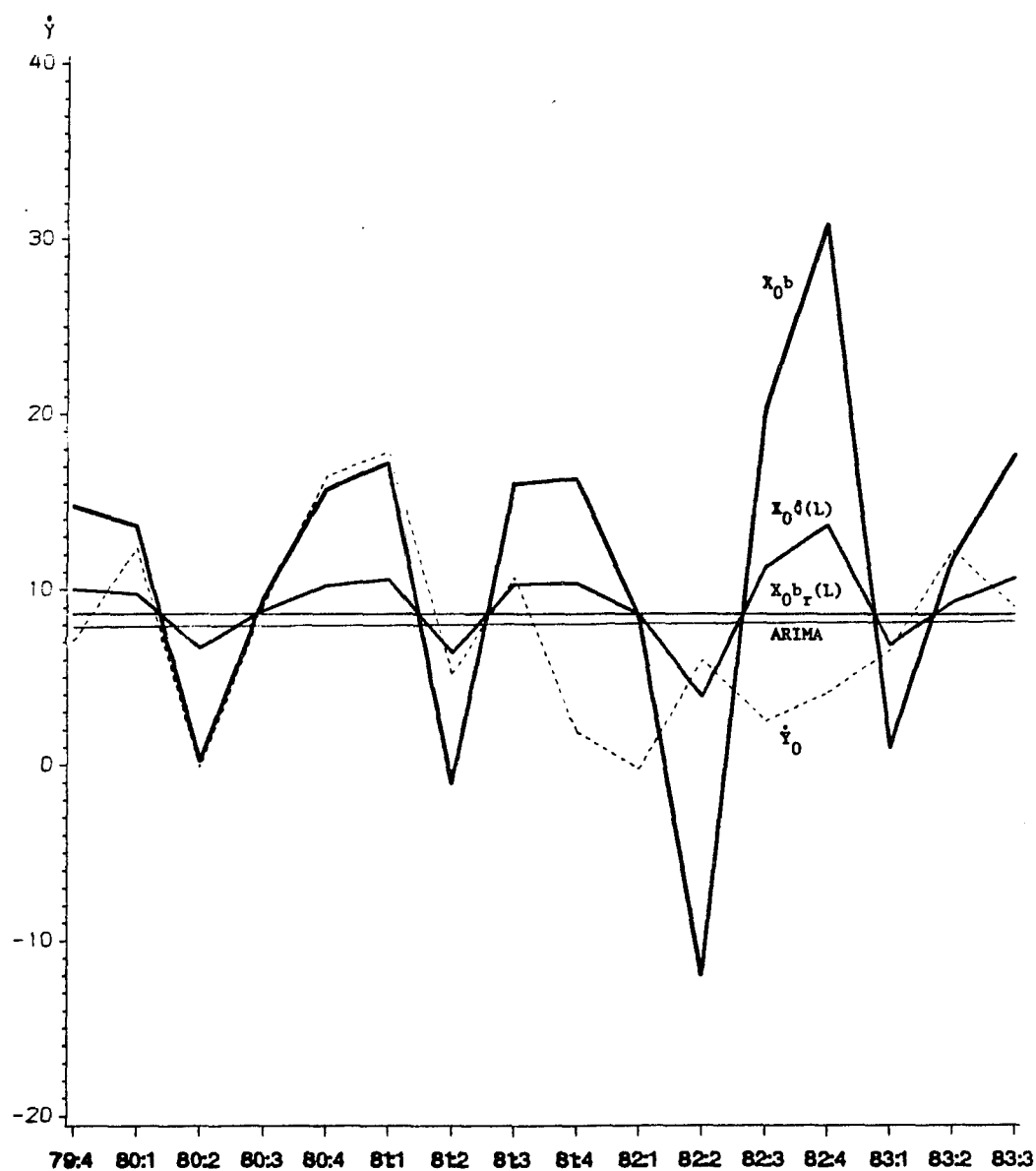


FIGURE 4.1

USING LS, LINDLEY, RLS(L), AND ARIMA
 FORECAST PERIOD: 1979:4-1983:3

FORECASTED VS. ACTUAL VALUES OF GNP GROWTH

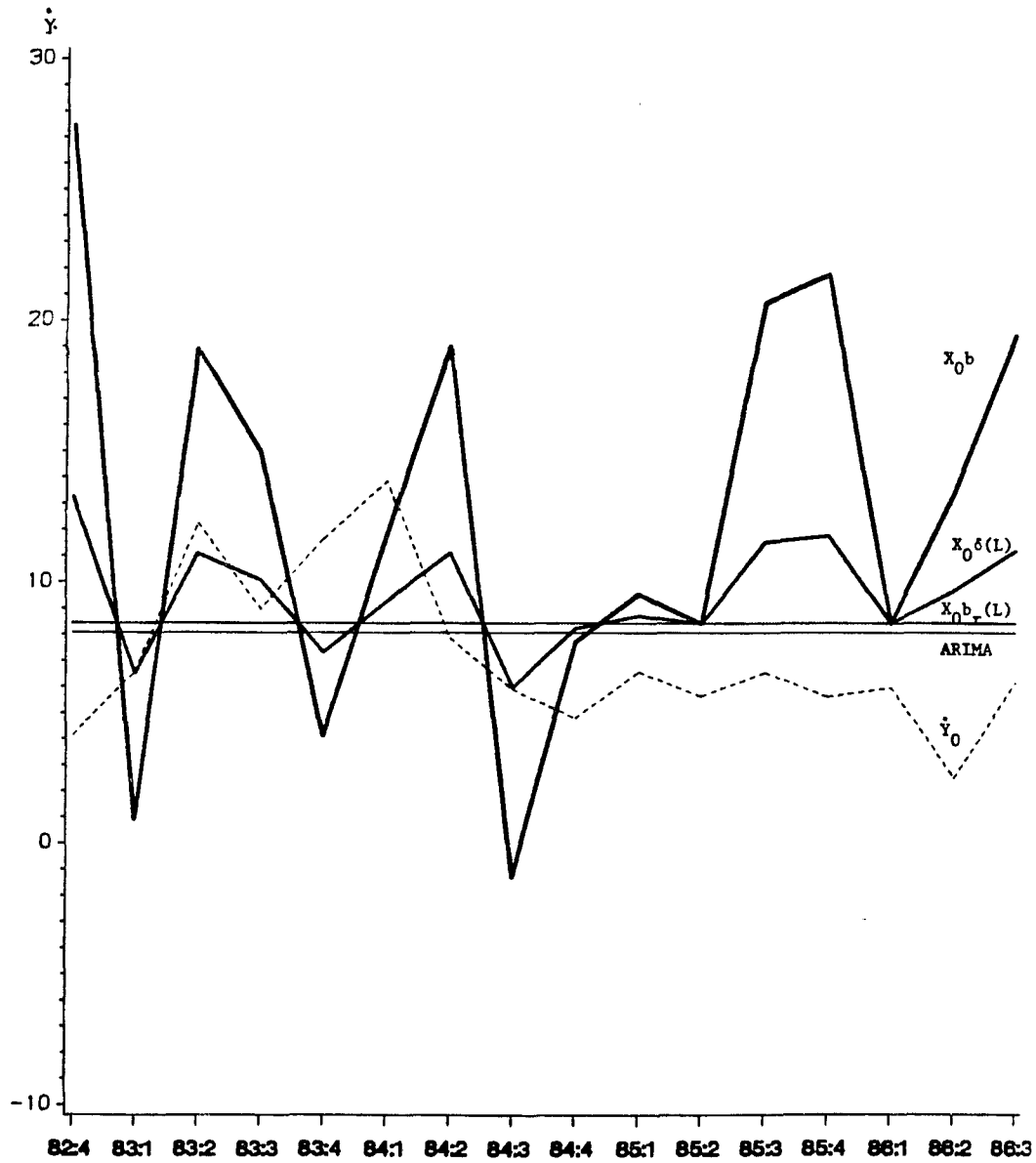


FIGURE 4.2

USING LS, LINDLEY, RLS(L), AND ARIMA
FORECAST PERIOD: 1983:4-1986:3

Appendix 4.1

Collinearity

In many instances, near linear dependencies among the included regressor variables can be a problem. This is especially true if many lagged variables are to be included in the model. When a significant degree of multicollinearity exists, researchers often resort to the use of uncertain prior information as a means of obtaining more efficient parameter estimates and forecasts. A common way to use prior information is to assume that the lag weights g_i and m_i lie along $p \leq 1$ and $q \leq k$ order polynomials, respectively [Batten and Thornton (1983), (1984) and Schmidt and Waud (1973)]. Although the imposition of correct restrictions leads to unbiased, consistent parameter estimates and more powerful hypothesis tests than ones based on the ordinary least squares estimator, the imposition of incorrect restrictions will lead to biased, inconsistent estimates and invalid tests. In terms of weighted quadratic risk, it is uncertain whether efficient biased forecasts will be better or worse than unbiased OLS forecasts.

The extent of the collinearity problem can be gauged by looking at the characteristic roots of the regressor cross product matrix, $S=X'X$. The characteristic roots measure squared deviations of the data along the axes of a new orthogonal basis defined by the corresponding characteristic vectors of S . Using this fact, Belsley,

Kuh, and Welsch (1980) have suggested calculating the condition number

$$c_i = (\lambda_1 / \lambda_i)^{1/2} \quad i=1,2,\dots,K$$

where $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_K$ are the characteristic roots of S_s , which is the scaled regressor cross product matrix. The number c_i is a measure of relative variation in the data; as the magnitude of c_i increases, one is able to conclude that in the new basis the data are becoming less variable in the direction of the i^{th} axis (relative to their variability along the major axis).

Characteristic roots of S_s for centered and uncentered data for each of the samples considered are reported below in Tables A.4.1.a and A.4.1.b. It is worth noting that for the 1962:2-1979:3 sample period, the two largest condition numbers for the centered data are 6.90 and 9.75, Belsley, Kuh, and Welch (1980) interpret this as indicating a "weak to moderate degree" of multicollinearity. For the uncentered data, the two largest condition numbers are 28.77 and 42.58. Based on uncentered data, one concludes that a "moderate to strong degree" of multicollinearity exists.

For the 1962:2-1982:3 sample period, the largest condition numbers are 4.35 and 5.12 for the centered data and 17.17 and 20.51 for the uncentered data. The degree of multicollinearity in the longer sample period does not appear to be as extreme as that in the 1962:2-1979:3 sample period.

Although the question of which set of condition numbers (centered or uncentered) to use as a collinearity diagnostic has been addressed, no consensus exists. See Belsley, Kuh, and Welsch (1980), Belsley (1984), and Snee and Marquardt (1984) for commentary on this debate. For the discussion below, it suffices to say that by either measure, the degree of multicollinearity appears to be a problem.

The presence of multicollinearity need not pose a serious threat to forecast accuracy. If the linear relationships among the out-of-sample regressors is similar to that existing among the in-sample regressors, forecast accuracy need not be impaired; this is the rationale for using assumption A1.3. Therefore, an attempt should be made to determine whether this is in fact the case. If not, then one can expect forecast accuracy to suffer.

Table A.4.1.a

Characteristic Roots and Condition Numbers
for Centered and Uncentered Data
(Scaled)

St. Louis Equation 1962:2 to 1982:3

	Centered Data		Uncentered Data	
	c_i	λ_i	c_i	λ_i
Row 1	1.000000	3.837856	1.000000	16.578040
Row 2	1.314515	2.221047	3.622114	1.263598
Row 3	1.366895	2.054086	3.779911	1.160299
Row 4	1.448942	1.828045	4.637689	0.770779
Row 5	1.485072	1.740177	4.806788	0.717502
Row 6	1.679341	1.360852	5.030211	0.655180
Row 7	1.807267	1.175016	5.256396	0.600007
Row 8	1.819505	1.159263	5.434879	0.561246
Row 9	1.835979	1.138553	5.487911	0.550451
Row 10	1.966444	0.992488	5.604323	0.527821
Row 11	2.041077	0.921234	5.893283	0.477329
Row 12	2.087096	0.881056	5.944229	0.469182
Row 13	2.173481	0.812412	6.163371	0.436411
Row 14	2.200549	0.792549	6.225853	0.427696
Row 15	2.271683	0.743692	7.826078	0.270672
Row 16	2.353759	0.692731	8.260680	0.242941
Row 17	2.635538	0.552522	9.156222	0.197742
Row 18	2.747416	0.508439	9.705329	0.176000
Row 19	2.862862	0.468260	10.406330	0.153086
Row 20	2.937855	0.444659	10.587540	0.147891
Row 21	3.153291	0.385976	11.019130	0.136533
Row 22	3.369889	0.337954	11.852200	0.118014
Row 23	3.509687	0.311567	12.362640	0.108470
Row 24	3.632285	0.290890	13.057950	0.097226
Row 25	4.357368	0.202134	16.584220	0.060275
Row 26	5.117693	0.146534	17.178580	0.056177
Row 27	-	-	20.506000	0.039424

Table A.4.1.b

Characteristic Roots and Condition Numbers
for Centered and Uncentered Data
(Scaled)

St. Louis Equation 1962:2 to 1979:3

	Centered Data		Uncentered Data	
	c_i	λ_i	c_i	λ_i
Row 1	1.000000	4.801358	1.000000	16.433900
Row 2	1.415937	2.394838	3.237719	1.567698
Row 3	1.495328	2.147294	3.427346	1.399023
Row 4	1.677173	1.706901	4.702811	0.743063
Row 5	1.769607	1.533240	4.763401	0.724279
Row 6	1.870730	1.371961	5.009032	0.654987
Row 7	1.968782	1.238708	5.126708	0.625264
Row 8	2.071339	1.119082	5.201311	0.607456
Row 9	2.125427	1.062849	5.565213	0.530612
Row 10	2.174544	1.015378	5.604882	0.523127
Row 11	2.229204	0.966194	5.839015	0.482016
Row 12	2.318665	0.893075	5.941391	0.465548
Row 13	2.416090	0.822503	5.998156	0.456778
Row 14	2.495982	0.770692	6.218090	0.425037
Row 15	2.565961	0.729229	7.467493	0.294707
Row 16	2.668190	0.674420	8.357008	0.235309
Row 17	2.737620	0.640645	9.352274	0.187891
Row 18	3.194019	0.470640	9.859245	0.169064
Row 19	3.407106	0.413616	10.967370	0.136626
Row 20	3.732330	0.344670	13.923170	0.084774
Row 21	4.070420	0.289792	15.596680	0.067557
Row 22	5.361359	0.167037	16.824420	0.058057
Row 23	5.666726	0.149520	20.358890	0.039649
Row 24	6.198808	0.124953	22.748820	0.031755
Row 25	6.900138	0.100840	24.723340	0.026886
Row 26	9.744804	0.050561	28.768020	0.019857
Row 27	-	-	42.576650	0.009065

Appendix 4.2

Assessment of Similarity of In-Sample and Out-of-Sample Regressor Matrices

In the absence of any well-developed procedure for measuring differences between in-sample and out-of-sample regressor matrices, conclusions from this section are quite tentative. Nevertheless, using the framework developed by Hill and Fomby (1986), an attempt is made to measure the extent to which A1.3 is violated for the model.

Using the geometry of principal components, Hill and Fomby (1986) develop several useful definitions in their discussion of multivariate data extrapolation. In effect, they argue that the in-sample regressor matrix X is similar to the out-of-sample regressor matrix X_0 if they have the same size, location, shape, and orientation in the regressor space.

To make this more concrete, let \bar{X} and \bar{X}_0 be $K \times 1$ vectors of the means of in-sample and out-of-sample regressors, respectively; let V and V_0 be the matrices of characteristic vectors of $X'X$ and $X_0'X_0$, respectively; and, let Λ and Λ_0 be diagonal matrices whose nonzero elements are the associated characteristic roots. In- and out-of-sample regressors are defined to be mean equivalent (i.e., $\bar{X} = \bar{X}_0$) if

$$d_1 = [(\bar{X} - \bar{X}_0)'(\bar{X} - \bar{X}_0)]^{1/2} = 0.$$

The regressor matrices are defined to be rotationally equivalent (i.e., have the same orientation in the data space) if $V = V_0$ and variationally equivalent (i.e., have the

same shape and size) if $\Lambda = \Lambda_0$. Finally, X and X_0 are said to be rotationally and variationally equivalent if $V\Lambda^{-\frac{1}{2}} = V_0\Lambda_0^{-\frac{1}{2}}$.

Using the data for the St. Louis equation, an attempt is made to measure the differences of mean, rotation, and variation between $X'X$ to $X_0'X_0$. The in-sample regressor matrix consists of observations from 1962:2-1979:3 and X_0 contains observations taken from 1979:4-1986:4. If the number of observations in the post-sample regressor matrix N is exceeded by the number of regressors K , then one is assured of having at least $K-N$ zero characteristic roots. In this instance, the two ellipsoids cannot be considered variationally or rotationally equivalent. Therefore, the period describe above is the only one considered.

Several tentative measures of ellipsoidal differences are given below. For these measures, the data were centered about the means.

Mean Difference

$$d_1 = [(\bar{X} - \bar{X}_0)'(\bar{X} - \bar{X}_0)]^{1/2} = 19.5$$

Rotational Differences

$$d_2 = |V_0'V - I| = 2.5 \text{ E-15}$$

$$d_3 = [\text{trace}(V_0'V - I)]^{1/2} = 7.43$$

Variational Differences

$$d_4 = [\text{trace}(\Lambda^{-\frac{1}{2}}\Lambda_0^{\frac{1}{2}} - I)]^{\frac{1}{2}} = 1.02$$

Rotational and Variational Differences

$$d_5 = |V_0'V\Lambda^{-\frac{1}{2}}\Lambda_0^{\frac{1}{2}} - I| = .109$$

$$d_6 = [\text{trace}(V_0'V\Lambda^{-\frac{1}{2}}\Lambda_0^{\frac{1}{2}} - I)]^{\frac{1}{2}} = 7.56$$

The problem with each of these measures of ellipsoidal differences is that no scale of reference has been found which permits one to tell whether a certain outcome is large or small. Thus, when it is computed that the Euclidean distance from the center of X to the center of X_0 is 19.5 very little else can be said.

The problems of scale notwithstanding, differences in mean, rotation, and variation are apparent. Perhaps it is safe to conclude that A1.3 does not hold strictly, and that in-sample and out-of-sample regressor matrices appear to be moderately dissimilar in the sense described above.

The characteristic roots for the (unscaled) regressor cross product matrices $X'X$ and $X_0'X_0$ are also given in Tables A.4.2.a and A.4.2.b. Notice that the characteristic roots of the in-sample and out-of-sample regressor cross product matrices are roughly similar in value except for the 13th and the 26th rows. The largest proportional difference occurs in row 26 where the smallest characteristic root of $X'X$ is over three times as large as the smallest root of $X_0'X_0$.

Table A.4.2.a

Characteristic Roots for Centered
In-Sample and Out-of-Sample Data
(Not Scaled)

	In-Sample	Out-of-Sample	Ratio
Row 1	11950.25	10975.87	1.088775
Row 2	11154.95	10964.16	1.017402
Row 3	5829.92	9254.98	0.629922
Row 4	5512.48	7976.79	0.691064
Row 5	5019.29	5917.07	0.848272
Row 6	4888.57	4671.19	1.046536
Row 7	4761.58	4153.20	1.146485
Row 8	3944.55	3685.95	1.070159
Row 9	3746.98	3626.39	1.033255
Row 10	3565.37	2174.27	1.639801
Row 11	3462.01	2126.03	1.628388
Row 12	3087.78	1992.09	1.550015
Row 13	2778.71	1474.45	1.884569
Row 14	1911.40	824.79	2.317440
Row 15	628.37	704.88	0.891460
Row 16	549.44	689.22	0.797198
Row 17	454.31	599.07	0.758355
Row 18	420.02	463.85	0.905510
Row 19	303.48	398.95	0.760707
Row 20	186.15	355.06	0.524269
Row 21	145.63	164.97	0.882758
Row 22	79.63	147.06	0.541516
Row 23	70.75	111.11	0.636784
Row 24	59.27	71.08	0.833816
Row 25	45.67	47.40	0.963549
Row 26	20.63	6.56	3.140870

Table A.4.2.b

Characteristic Roots for Centered
In-Sample and Out-of-Sample Data
(Scaled)

	In-Sample	Out-of-Sample	Ratio
Row 1	4.801358	2.560916	1.874859
Row 2	2.394838	2.446527	0.978872
Row 3	2.147294	2.294197	0.935967
Row 4	1.706901	2.231507	0.764909
Row 5	1.533240	2.141877	0.715839
Row 6	1.371961	1.806399	0.759500
Row 7	1.238708	1.601359	0.773535
Row 8	1.119082	1.450051	0.771753
Row 9	1.062849	1.269283	0.837362
Row 10	1.015378	1.164445	0.871985
Row 11	0.966194	1.086788	0.889036
Row 12	0.893075	0.900076	0.992221
Row 13	0.822503	0.866937	0.948746
Row 14	0.770692	0.790891	0.974460
Row 15	0.729229	0.755175	0.965642
Row 16	0.674420	0.524381	1.286124
Row 17	0.640645	0.518086	1.236561
Row 18	0.470640	0.448954	1.048302
Row 19	0.413611	0.445320	0.928795
Row 20	0.344670	0.305117	1.129633
Row 21	0.289792	0.136738	2.119315
Row 22	0.167037	0.091296	1.829614
Row 23	0.149520	0.079582	1.878815
Row 24	0.124953	0.044187	2.827771
Row 25	0.100843	0.032794	3.075030
Row 26	0.050561	0.007111	7.109691

Appendix 4.3

Regression Diagnostics

Verification of assumption A1.2 boils down to determining whether $\epsilon \sim N(0, \sigma^2 I)$ based on estimates of the residuals. Below, tests for nonnormality, heteroscedasticity, and autocorrelation are considered and results presented.

A.4.3.1 Nonnormality

To detect possible departures of the residual vector e from the normality assumption, several tests based on least squares residuals from (4.3.1) are performed. These tests include the well-known Kolmogorov-Smirnov D test (Pearson and Hartley, 1966), the Anderson-Darling A^2 test (Stephens, 1974), and tests based on the measures of skewness \hat{b}_1 and kurtosis \hat{b}_2 [see White and MacDonald, (1980); and, Pearson and Hartley, (1966) for tables].

The omnibus tests based on the Kolmogorov-Smirnov D^+ , D^- , and D and the Anderson-Darling A^2 are developed below. First, let the order statistics of the least squares residuals be denoted $e_1 \leq e_2 \leq \dots \leq e_T$. Using these, calculate $w_i = (e_i - \bar{e})/\tilde{\sigma}$, where \bar{e} is the mean of e_i $i=1, 2, \dots, T$ and $\tilde{\sigma} = [\sum e_i^2 / T]^{1/2}$. Now, let $z_i = \Phi(w_i)$ be the standard normal integral evaluated at the argument w_i . The Kolmogorov statistics D^+ , D^- , and D are defined to be

$$\begin{aligned} D^+ &= \max_{1 \leq i \leq T} [i/t - z_i] \\ D^- &= \max_{1 \leq i \leq T} [z_i - (i-1)/T] \\ D &= \max(D^+, D^-) \end{aligned}$$

According to Stephens (1974) the statistic $T(D)$, for use when the mean μ and variance σ^2 of the variable of interest are unknown (Case III), is

$$T(D) = D(\sqrt{T} - .01 + .85/\sqrt{T}).$$

Stephens notes, however, that $T(D)$ is low in power relative to the Anderson-Darling statistic A^2 which is defined to be

$$A^2 = - \left\{ \sum_{i=1}^T (z_i - 1) [\ln(z_i) + \ln(1 - z_{T+1-i})] \right\} / T - T$$

The appropriate modification for use when μ and σ^2 are unknown (Case III) is

$$T(A^2) = A^2 [1 + 4/T - 25/T^2].$$

An argument can be made that since by assumption $E(e)=0$, it is appropriate to devise a test which is sensitive to departures in the data from zero mean. To verify whether $e \sim N(0, \sigma^2 I)$, one is ultimately interested in the location as well as its shape of the distribution. When μ is known Stephens (1974) suggests using $w_i^* = (e_i - \mu) / \sigma$ in place of w_i in D and A^2 above. This situation is referred to by Stephens as Case II for which the appropriate transformations $T(D)$ and $T(A^2)$ are

$$T^*(D) = \sqrt{T} D$$

$$T^*(A^2) = A^2 \quad T \geq 5.$$

The test for nonnormality of residuals in regression analysis, Stephens (1974) argues, should be applied to transformed linear combinations of the OLS residuals which are theoretically distributed independently normal with zero mean. Therefore an appropriate test of $H_0: e \sim N(0, \sigma^2)$

vs. $H_a: e \sim N(\mu, \sigma^2 I)$ can be conducted using BLUS residuals. The results of Case II and Case III tests for nonnormality are summarized below in Tables A.4.3.a and A.4.3.b.

The statistics associated with the omnibus tests D and A^2 for Cases II and III yield p-values greater than .15 for both sample periods. The p-values from the tests based on the Case II assumption of known mean and unknown variance which use the BLUS residuals suggest that the 62:2-82:3 residuals are more likely to be independent normal with zero mean than those of the 62:2-79:2 sample. The large difference in test statistics is due to the fact that the mean of the BLUS residuals for the 62:2-79:3 sample is -.68 verses a mean of -.056 for the longer sample period.

Case III p-values indicate that when the mean of the empirical distribution function (EDF) is estimated, the 1962:2-1982:3 residuals appear more likely to be normal than those of the 62:2-79:3 sample period. Though normality cannot be rejected at usual levels of significance (5% or 10%), a comparison of the Case II p-values would seem to indicate that the 62:2-79:3 specification appears to be more likely to violate the assumption of having zero mean.

The statistics $\sqrt{\hat{b}_1}$ and \hat{b}_2 are measures of skewness and kurtosis. The ones calculated in this study are discussed in White and MacDonald (1980) where they are defined to be

$$\sqrt{\hat{b}_1} = T^{-1} \sum e_i^3 / (\hat{\sigma}^2)^{3/2}$$

and

$$\hat{b}_2 = T^{-1} \sum e_i^4 / (\hat{\sigma}^2)^2.$$

Critical values for $\sqrt{\hat{b}_1}$ and \hat{b}_2 are found in Pearson and Hartley (1966). The computed values of these statistics and their approximate 90% confidence intervals are given in Table A.4.3.c. The values of the statistic $\sqrt{\hat{b}_1}$ suggests that neither sample is badly skewed. In addition, \hat{b}_2 falls within the 90% confidence interval for both samples; however, there appears to be some evidence that the distribution of the 1962:2-1979:3 residuals is leptokurtotic (high peak).

Table A.4.3.a
Tests for Nonnormality[†]
Case II, σ^2 unknown

Sample	Statistic	Computed Value	Percentage Points ^{††}			
			15	10	5	2.5
62:2-82:3	$T^*(D)$.3295	1.082	1.171	1.311	1.433
62:2-82:3	$T^*(A^2)$.1340		1.760	2.323	2.904
62:2-79:3	$T^*(D)$	1.1557	1.078	1.168	1.305	1.428
62:2-79:3	$T^*(A^2)$	1.3269		1.760	2.323	2.904

Table A.4.3.b
Tests for Nonnormality^{†††}
Case III, μ , σ^2 unknown

Sample	Statistic	Computed Value	Percentage Points ^{††}			
			15	10	5	2.5
62:2-82:3	$T(D)$.5089	.775	.819	.895	.955
62:2-82:3	$T(A^2)$.1704	.576	.656	.787	.918
62:2-79:3	$T(D)$.5694	.775	.819	.895	.955
62:2-79:3	$T(A^2)$.3905	.576	.656	.787	.918

Table A.4.3.c
Nonnormality Tests

Sample	Statistic	Computed value	90% ^{††††}
			Confidence Interval
62:2-82:3	$\hat{\sqrt{b}}_1$	-0.0959	[-0.427, 0.427]
62:2-82:3	\hat{b}_2	2.836	[2.29, 3.87]
62:2-79:3	$\hat{\sqrt{b}}_1$	-0.115	[-0.459, 0.459]
62:2-79:3	\hat{b}_2	3.78	[2.25, 3.89]

[†]Tests to detect nonnormality and nonzero mean of e using OLS residuals of model 3.1 ($l=k=12$) transformed to BLUS residuals.

^{††}Tables from Stephens (1974).

^{†††}Tests to detect nonnormality of residuals for model 3.1 ($l=k=12$) using least squares residuals.

^{††††}Tests based on measures of skewness and kurtosis of residuals for model 3.1 ($l=k=12$) using least squares residuals. Interpolation used to get confidence intervals, see White and MacDonald for tables (1984).

A.4.3.2 Autocorrelation

The usual test for first order autocorrelation [AR(1)] is the Durbin-Watson bounds test. The difficulty of performing this test, which uses the least squares residuals of (4.3.1), arises from the fact that upper and lower bounds have not been tabled for $K = 27$. In addition, the calculated values of the Durbin Watson statistic (2.14 for 62:2-82:3 and 2.23 for 62:2-79:3) are likely fall within the inconclusive region, which becomes quite large as the number of regressors increases. Simple extrapolation using the tables for the largest value of K available ($K=21$) indicates this to be so.

Another procedure for testing the presence of AR(1) in least squares residuals is the BLUS test of Abrahamse and Koerts (1968) which uses the statistic

$$Q^* = \frac{\sum_{t=1}^{T-K} (e_t^* - e_{t-1}^*)^2}{\sum_{t=1}^{T-K} (e_t^* - \bar{e}^*)^2}$$

where e_t^* $t=1,2,\dots,T-K$ are the BLUS residuals (see Theil

(1971), pp. 202-206) and $\bar{e}^* = \sum_{t=1}^{T-K} e_t^* / (T-K)$. Unlike the

Durbin-Watson statistic, the significance points q^* , which have been tabled by Abrahamse and Koerts (1971), do not depend on X . The null hypothesis of no first-order autocorrelation is rejected if

$$Q^* \leq q^* \text{ (if } \rho > 0 \text{) or if } Q^* \geq 4 - q^* \text{ (if } \rho < 0 \text{)}.$$

The use of the BLUS test is complicated by having to choose an appropriate basis for the linear transformation

of OLS to BLUS residuals. Following the suggestion of A-K [(1971), pp. 948-949] and Theil [(1971), p. 217] the basis was chosen such that $\sum \lambda(i)$ $i=1,2,\dots,K$ is maximized, where $\lambda(i)$ is the i^{th} characteristic root of the matrix $X_0(X'X)^{-1}X_0$. The details of this procedure are found in the references cited above.

For the 1962:2-1982:3 sample period, the computed value of Q^* is 1.989. Using the tables in Abrahamse and Koerts (1971) the 10% significance point of Q^* such that $\Pr[Q^* < q^*]$ is 1.6587. Since 1.989 falls within the 80% confidence interval [1.6567, 2.3413], one would not reject the null hypothesis $H_0: \rho=0$ against all alternatives at the 20% level of significance. Likewise, for the 1962:2-1979:3 sample period, where $Q^*=1.948$ falls well within the 80% confidence interval [1.6154, 2.3486].

A.4.3.3 Heteroscedasticity

One hypothesis worth exploring with time series data is whether variance is constant over time. Three tests for detecting the assumption of constant variance were performed for (4.3.1). The first reported is the Goldfeld-Quandt (1965, 1972) test. For this test, p central observations are deleted and separate regressions are run for the two sample subsets. Then, an F-test is formed by using the ratio of sum-of-squared error functions from the two regressions. Table A.4.3.d summarizes the results of the Goldfeld-Quandt test for the two sample periods.

Although the sum-of-squared errors appears to be

greater in the latter subsamples (i.e., $sse2 > sse1$), the nominal marginal significance level of the F-statistics is never less than .34, a level well above that traditionally used to test hypotheses (i.e., .05 or .1). That is to say, at the 5% or 10% level of significance, the null hypothesis of homoscedasticity cannot be rejected on the basis of this test.

Table A.4.3.d
Goldfeld-Quandt Tests
Model (4.3.1)

Sample	p	SSE1	SSE2	F	P-Value
1962:2-1979:3 T=70, K=27	12	32.39	46.85	1.45 ($F_{2,2}$)	.41
1962:2-1982:3 T=82, K=27	20	38.36	59.48	1.55 ($F_{4,4}$)	.34

White's test (1980) is a general test used to detect misspecification of the linear form (4.3.1) as well as heteroscedastic error variance. In general, one would like to jointly test these propositions; however, the dimension of the regression problem (i.e., $K=27$) is too large for the number of available observations and White's test cannot be performed. Therefore, the Breush-Pagan procedure is pursued as a computationally efficient alternative. In addition, results of Ramsey's RESET test for misspecification are reported.

Although the Breush-Pagan (1979, 1982) test does not require specification of the functional form of the heteroscedasticity, one must nevertheless make some

conjecture about the set of possible variables to which the variance is related. Again, the proposition is entertained that error variance changes over time. That is $\sigma_i^2 = z_i' \alpha$ $i=1,2,\dots,T$ where z_i is an $L \times 1$ vector of variables affecting the i^{th} error variance. Let $q = (\sigma_1^2, \dots, \sigma_T^2)'$ be the $T \times 1$ vector of error variances and consider two matrices, Z_a and Z_b , of variables suspected of influencing error variance.

$$Z_a' = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 2 & \dots & T \end{bmatrix} \quad Z_b' = \begin{bmatrix} 1 & \dots & 1 & 1 & \dots & 1 & 1 & \dots & 1 \\ 1 & \dots & 1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 & 1 & \dots & 1 \end{bmatrix}$$

The matrix Z_a is used under the assumption that error variance increases or decreases over time. The matrix Z_b contains an intercept term and 2 dummy variables which permits one to break the sample into 3 sections and test the equivalence of variance across the three sub-samples.

Computationally, Koenker's (1981) studentized version of Breush and Pagan's Lagrange multiplier statistic is calculated. Let $Z = Z_a$ or Z_b and

$$n = [\hat{q}' Z (Z' Z)^{-1} Z' \hat{q}] / \hat{q}' \hat{q}$$

where \hat{q} has t^{th} element equal to $\hat{q}_t = \hat{e}_t^2 - \hat{\sigma}^2$, $\hat{e}_t = y_t - x_t' b$, $t=1,2,\dots,T$, and $\hat{\sigma}^2 = \hat{e}' \hat{e} / T$. Under the null hypothesis of homoscedasticity (due to the variables in Z), $n \sim \chi^2(L-1)$. The results of tests based on Z_a and Z_b above are given in Table A.4.3.e below.

In conclusion, there appears to be little evidence to suggest that the error variance is not homoscedastic.

Table A.4.3.e
Breush-Pagan Tests
Model (4.3.1)

Sample Period	Hypothesis (Z_a or Z_b)	Test Statistic	P-Value
1962:2-1979:3	Z_a	.006	.93
1962:2-1979:3	Z_b	.097	.95
1962:2-1982:3	Z_a	.23	.63
1962:2-1982:3	Z_b	.50	.78

Note: The test statistics associated with use of Z_a and Z_b are distributed $\chi^2(1)$ and $\chi^2(2)$, respectively.

Finally, Ramsey's RESET test for misspecification was performed. The regression specification error test (RESET) uses BLUS residuals to test

$$H_0: e^* \sim N(0, I_{T-K}) \text{ vs. } H_a: e^* \sim N(A'\epsilon, \sigma^2 I)$$

where $A'\epsilon \neq 0$. Thus, RESET is a test for the presence of specification bias. For details of the test see Ramsey (1969, 1974). This test was conducted using the third order of approximation producing an F-statistic of 1.42 ($F_{4,50}$) with a p-value of approximately .23 for the 62:2-82:3 sample period and an F-statistic of 1.26 ($F_{4,38}$) with p-value .30 for the 62:2-79:3 sample period.

In summary, assumption A1.2 that $e \sim N(0, \sigma^2 I)$ is sustained for the statistical model (4.3.1) for the St. Louis equation over each of the sample periods considered.

Chapter 5

Improved Confidence Intervals and Ellipsoids for the Linear Regression Model

- 5.1 Interval Estimation and Hypothesis Testing Using Biased Estimators
- 5.2 Statistical Model and Data Generation Process
- 5.3 Estimators
- 5.4 The Bootstrap
 - 5.4.1 Bootstrap Estimates
 - 5.4.2 Percentile Confidence Intervals
- 5.5 Confidence Ellipsoids
 - 5.5.1 Alternative Specifications of the Quadratic Form
 - 5.5.2 Estimating Covariance
 - 5.5.3 Estimating Ellipsoids
 - 5.5.3 Obtaining Critical Points
 - 5.5.5 Summary
- 5.6 Results
 - 5.6.1 Intervals
 - 5.6.2 Ellipsoids
- 5.7 Conclusion

Chapter 5

Improved Confidence Intervals and Ellipsoids for the Linear Regression Model

5.1 Interval Estimation and Hypothesis Testing Using Biased Estimators

If Stein-rule estimation is to ever gain widespread acceptance among applied economic researchers, an acceptable measure of precision must be developed. Although there are exceptions (ridge estimator, pretest estimator, etc.), knowledge of the sampling distribution of an estimator is an important pre-condition for its use. The exact covariance matrix of the Stein-rule estimator is known [Judge and Bock, Section 8.9, (1978)], but the formula contains unknown population parameters. If one attempts to replace the unknown parameters with estimates, the sampling distribution of the usual quadratic forms are no longer chi-square or Snedecor's F. Consequently, interval estimates and hypothesis tests cannot be formulated in the usual fashion.

Phillips (1984) has been able to show that the exact distributional properties of the Stein-rule are, as he puts it, "well within reach." Specifically, he provides a formula for the probability density function of the James-Stein (1961) estimator and deduces moment formulae directly from this general result. Unfortunately, Phillip's results rely on the use of advanced analysis (Weyl fractional calculus) and cannot be readily implemented at this time.

Several alternatives have been pursued. One is to approximate the sampling distribution of the statistics of

interest by an asymptotic expansion. Ullah (1982) and Ullah, Carter, and Srivastava (1984) use an Edgeworth-type asymptotic expansion to approximate the multivariate and marginal sampling distributions for a class of biased estimators which includes those of the Stein-family and the corresponding overall F-statistic. Ohtani (1986) derives the distribution of an improved F-ratio [Ullah, Carter, and Srivastava, (1984)] obtained by using the James-Stein estimator in place of the OLS estimator and shows that the test based on the improved F-ratio for the null hypothesis that all regression coefficients are zero can be performed using the F-distribution. However, he also concludes that the power of this test is lower than that of the test given by the usual F-ratio.

Another line of research pursues Stein's (1962) conjecture that it is possible to derive improved confidence sets for the mean of a multivariate normal distribution. An improved confidence set is one with higher coverage probability and of no greater volume than the usual one--a sphere or ellipsoid of fixed volume centered at the sample mean. Brown (1966) and Joshi (1967) independently demonstrated the existence of improved confidence sets when the multivariate normal random vector contains at least 3 elements. Olshen (1977) simulated the coverage probabilities of Joshi's estimator and found that the improvements could be substantial under certain parameterizations.

Using empirical-Bayes techniques, Morris (1977, 1983) has shown that coverage probabilities of certain generalized-Bayes estimators are quite good. Berger (1980a) has constructed Bayesian confidence ellipsoids by considering the posterior covariance matrix. These ellipsoids are shown to have higher coverage probability over a significant portion of the parameter space and to be of uniformly smaller volume. Hwang and Casella (1982) have devised an explicit procedure for uniformly increasing coverage probability by centering the usual confidence set at the positive-part James-Stein estimator. Their result holds provided that the multivariate normal random vector has at least 4 elements. In addition, Hwang and Casella show that the possible improvement can be quite substantial.

The Berger (1980a) and Hwang and Casella (1982) estimators are limited to cases where the confidence sets are spherical. Hill and Fomby (1986) have examined the coverage probability and volume of Berger's estimator relative to OLS under a range of conditions commonly found in econometric practice. Surprisingly, they find Berger's estimator to be quite robust to various degrees of multicollinearity.

Most of the research on this topic has pursued the Bayesian confidence set approach. Many economists, however, are reluctant to embrace Bayesian and empirical Bayesian methods. If progress is to be made toward finding

acceptable measures of precision, then an alternative must be found which not only yields exact or approximate tests of a given size with adequate power, but which is also reasonably easy to perform.

One possibility is to use Efron's bootstrap [Efron, (1979, 1981, 1982, 1986); Freedman, (1981)] which is a general procedure for measuring the variability of a statistic having an unknown sampling distribution. In essence, bootstrapping permits one to approximate the sampling distribution of a statistic by replacing the unknown distribution function with the empirical distribution of the data and then resampling randomly to obtain a Monte Carlo distribution of the resulting random variable. Chi and Judge (1985) have compared confidence intervals for the James-Stein estimator based on bootstrap resampling with those derived via empirical Bayes estimation under the assumption that σ^2 is known.

In this chapter, bootstrap confidence intervals are constructed using Efron's (1979, 1982) percentile method and compared to the bootstrap method employed by Chi and Judge (1985) which is based on the approximate normality of the James-Stein rule estimator. Then, percentile bootstrap confidence ellipsoids centered at the James-Stein estimator are proposed and estimated. To measure the success of the bootstrap method, confidence intervals and ellipsoids centered at the James-Stein estimator are compared to those centered at least squares.

The chapter contains 6 sections in addition to the introduction. In section 5.2 the statistical model and the data generation process for the Monte Carlo is introduced. In section 5.3 the estimators of the parameters of the classical normal linear regression model are presented and the properties of the resulting estimates generated in the Monte Carlo are discussed. In section 5.4 the bootstrap is explored and in 5.5 confidence intervals are defined. In section 5.6 the coverage probabilities and sizes of the bootstrap confidence procedures are examined and in section 5.7 conclusions are drawn.

5.2 Statistical Model and Data Generation Process

In this chapter, we consider the size and coverage probability of several confidence intervals and ellipsoids centered at both unbiased and biased estimators of the classical normal linear regression model (CNLRM). The CNLRM is denoted

$$y = X\beta + e \quad e \sim N(0, \sigma^2 I) \quad (5.2.1)$$

where y is a $T \times 1$ vector of observable random variables, X is a known $T \times K$ nonstochastic design matrix of rank K , β is a $K \times 1$ vector of unknown parameters, and e is a $T \times 1$ vector of unobservable normally and independently identically distributed random variables having zero mean and finite variance.

In the Monte Carlo experiment the orthonormal linear regression model with 8 regressors and no intercept is used (i.e., X is chosen such that $X'X = I_8$). The use of an

orthonormal design implies that regressors are mutually independent; hence, collinearity is not a problem. The orthonormal model corresponds to the "model of the mean" of a multivariate normal population which is widely studied in statistics. In addition, it is important to learn how new types of confidence procedures perform under circumstances like these in order to measure the impact of departures from the ideal. It remains to be seen what impact various degrees of multicollinearity will have on our results.

We draw $M=400$ random samples of size 30 from the $N(0,1)$ density using SAS (1986) RANNOR. Ten values of the parameter vector β are used in the simulation and chosen so that $\beta=cL$, where L is an 8×1 vector of ones,

$$c = \{R^2 T \sigma^2 / (1-R^2) L' L\}^{1/2} \quad (5.2.2)$$

$\sigma^2=1$, and population goodness-of-fit $R^2=[0.00001, 0.01, 0.025, 0.05, 0.075, 0.10, 0.25, 0.50, 0.75, 0.90]$. The same set of normal random deviates is used for each of the 10 parameter points implied by use of (5.2.2).

5.3 Estimators

The least squares (LS) and maximum likelihood estimator (MLE) of β in the classical normal linear regression model is

$$b = (X'X)^{-1} X'Y \sim N(\beta, \sigma^2 (X'X)^{-1}). \quad (5.3.1)$$

The estimator $\hat{\sigma}^2 = (Y - Xb)'(Y - Xb) / (T-K) = s^2 / (T-K)$ is the minimum variance unbiased estimator of σ^2 , $(T-K)\hat{\sigma}^2 / \sigma^2 \sim \chi^2_{T-K}$, and $\hat{\sigma}^2$ is statistically independent of b .

Let \tilde{b} be an arbitrary estimator of β and W be any

known, positive definite symmetric matrix. Weighted squared error loss is defined to be

$$L(\beta, \tilde{b}; W) = (\tilde{b} - \beta)' W (\tilde{b} - \beta) \quad (5.3.2)$$

The risk of using \tilde{b} to estimate the unknown vector β is obtained by considering the average loss over all samples

$$R(\beta, \tilde{b}; W) = E_Y[L(\beta, \tilde{b}; W)] = E[(\tilde{b} - \beta)' W (\tilde{b} - \beta)] \quad (5.3.3)$$

James and Stein (1961) have proposed an estimator of β which has risk no greater than the MLE under quadratic loss for all values of β ; hence, the MLE is inadmissible. The James-Stein estimator is

$$\delta(b) = (1 - c/u)b \quad (5.3.4)$$

where $c = a(T-K)/K$, $u = b'Sb/K\sigma^2 \sim F_{K, T-K, \lambda}$, $S = X'X$, $\lambda = \beta'S\beta/2\sigma^2$, and for minimaxity $K \geq 3$,

$$0 \leq a \leq [2/(T-K+2)][\eta_L^{-1} \text{tr}(WS^{-1}) - 2],$$

and η_L is the largest characteristic root of WS^{-1} .

The James-Stein estimator has mean

$$E[\delta(b)] = \beta + a(T-K)E[1/X_{K+2, \lambda}^2]\beta \quad (5.3.5)$$

and covariance matrix

$$E[(\delta - \beta)(\delta - \beta)'] = \sigma^2 S^{-1} - \sigma^2 S^{-1}[2a E(l_1) - a^2 E(l_1)^2] + \beta\beta' S^{-1}\{2a[E(l_1) - E(l_2)] + a^2[E(l_2)^2 - (E l_1)^2]\} \quad (5.3.6)$$

where $l_1 = X_{T-K}^2/X_{K+2, \lambda}^2$, $l_2 = X_{T-K}^2/X_{K+4, \lambda}^2$, and $\lambda = \beta'X'X\beta/2\sigma^2$.

For $\beta \neq 0$ it can be seen that the James-Stein estimator is biased in small samples and that its covariance matrix depends on the unknown location and scale parameters. In addition, the James-Stein estimator is nonlinear and

nonnormally distributed. This implies that hypothesis tests and confidence intervals of given size cannot be formed in the usual way, i.e., based on the Wald principle [Engle (1984)]. This inability to estimate confidence intervals detracts from the use of the Stein type estimators in most applied work.

As a practical matter in Monte Carlo studies, it is often useful to examine the sample moments of the statistics of interest and to compare these to theoretical values when known. The Monte Carlo sample moments of the estimators b and δ are examined in Tables 5.1 and 5.2 below. In Table 5.1, we report the mean, standard deviation, skewness and kurtosis of each element of b and δ using

$$\mu_{1i} = M^{-1} \sum_{m=1}^M \tilde{b}_{m,i}$$

$$\mu_{2i} = M^{-1} \sum_{m=1}^M (\tilde{b}_{m,i} - \mu_{1i})^2$$

$$\mu_{3i} = M^{-3/2} \sum_{m=1}^M (\tilde{b}_{m,i} - \mu_{1i})^3 / \left[\sum_{m=1}^M (\tilde{b}_{m,i} - \mu_{1i})^2 \right]^{3/2}$$

$$\mu_{4i} = M \left\{ \sum_{m=1}^M (\tilde{b}_{m,i} - \mu_{1i})^4 / \left[\sum_{m=1}^M (\tilde{b}_{m,i} - \mu_{1i})^2 \right]^2 \right\} - 3$$

where $\tilde{b}_{m,i}$ is the estimate of β_i from the m^{th} iteration of the Monte Carlo. Notice that 3 has been subtracted from the usual estimate of kurtosis in order to make comparisons with the normal density easier to interpret. Thus, if $\hat{b}_i \sim N(\beta_i, 1)$, then $\mu_{1i} \cong \beta_i$, $\mu_{2i} \cong 1$, $\mu_{3i} \cong 0$, and $\mu_{4i} \cong 0$.

From Table 5.1 it can be seen that least squares estimates are approximately equal to their expected values, their standard errors are approximately equal to one, and they appear to have a negligible degree of skewness and kurtosis. A histogram (which has been smoothed) depicting the sample distribution of b_1 over the 400 Monte Carlo iterations appears in Figure 5.1. Notice that it closely approximates the histogram obtained from 400 $N(0,1)$ deviates.

From Table 5.1 we also can see that James-Stein estimates are unbiased at the origin. As we expect, bias increases as we move away from the origin and then decreases for $R^2 > 0.5$. The standard errors of the James-Stein estimates from the Monte Carlo are quite accurate. They tend to be slightly underestimated for small values of R^2 and overestimated for larger values ($R^2 \geq 0.5$).

In Figure 5.2, notice how the shrinkage estimator concentrates point estimates in a very tight region near the true parameter point when $R^2 = .00001$. As we move away from the origin, the tightness relaxes and the standard error increases. Skewness at the origin does not appear to be significant, but it increases as R^2 approaches .25 and declines thereafter. The estimates of kurtosis indicate that near the origin the distribution is highly peaked (leptokurtotic); the peak falls and the tails get 'fatter' as R^2 increases. The smoothed histograms of δ_1 and b_1 are compared to one another for $R^2 = .00001$ in Figure 5.3.

Clearly, δ_1 is highly concentrated about 0 while least squares estimates are more highly dispersed. In Figure 5.4 smoothed histograms appear for δ_1 for several values of R^2 . Notice that as R^2 increases, the mean and dispersion of the James-Stein estimates increases.

In Table 5.2 we average the results of Table 5.1 across elements (i.e., $\bar{\mu}_p = \sum_{i=1}^K \mu_{pi} / K$ $p=1,2$) and report the difference between b and its expectation as well as the ratio of estimated to expected standard error. This yields overall information about how well the first two central moments of the Monte Carlo distribution approximates the expectations of b and δ and their standard errors.

From Table 5.2 it is apparent that in the Monte Carlo μ_1 and μ_2 conform with a high degree to their expectations. Note from the bottom portion of that table that in the Monte Carlo least squares is approximately unbiased (.017) and that the standard error of b is overestimated by about 0.3%. For the James-Stein estimator, bias is within 0.03 for each value of R^2 . Note however that overall, the standard error is underestimated by 3.8% near the origin and overestimated by a scant .1% for $R^2=.9$. The degrees of over- and underestimation of the standard error of the James-Stein estimator are very small in the Monte Carlo.

Finally, Kolmogorov's D statistics are reported in Table 5.3. The D statistic is used to detect departures from normality in the sample. The normality hypothesis (mean and variance unknown) is rejected when D is large.

Table 5.1
Summary Statistics
Monte Carlo

	R^2	.00001	.010	.025	.050	.075	.100	.250	.500	.750	.900
β_i		.0069	.218	.345	.487	.596	.689	1.089	1.541	1.887	2.067
b_1	μ_1	-0.006	0.204	0.331	0.474	0.583	0.676	1.076	1.528	1.874	2.054
	$\sqrt{\mu_2}$	0.967	0.967	0.967	0.967	0.967	0.967	0.967	0.967	0.967	0.967
	μ_3	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048
	μ_4	-0.249	-0.249	-0.249	-0.249	-0.249	-0.249	-0.249	-0.249	-0.249	-0.249
δ_1	μ_1	0.003	0.075	0.133	0.216	0.293	0.365	0.735	1.213	1.590	1.786
	$\sqrt{\mu_2}$	0.499	0.518	0.543	0.579	0.609	0.633	0.736	0.822	0.861	0.876
	μ_3	0.052	0.413	0.531	0.602	0.641	0.665	0.541	0.308	0.206	0.172
	μ_4	2.049	1.985	1.890	1.626	1.308	1.002	0.166	-0.115	-0.178	-0.194
b_2	μ_1	-0.022	0.188	0.315	0.458	0.567	0.660	1.060	1.512	1.858	2.038
	$\sqrt{\mu_2}$	1.027	1.027	1.027	1.027	1.027	1.027	1.027	1.027	1.027	1.027
	μ_3	-0.381	-0.381	-0.381	-0.381	-0.381	-0.381	-0.381	-0.381	-0.381	-0.381
	μ_4	0.850	0.850	0.850	0.850	0.850	0.850	0.850	0.850	0.850	0.850
δ_2	μ_1	-0.049	0.032	0.095	0.183	0.262	0.336	0.715	1.198	1.576	1.772
	$\sqrt{\mu_2}$	0.569	0.587	0.600	0.623	0.647	0.670	0.771	0.859	0.904	0.921
	μ_3	-0.928	-0.496	-0.340	-0.172	-0.036	0.054	0.119	-0.060	-0.169	-0.209
	μ_4	6.047	4.503	3.636	2.896	2.367	1.957	0.883	0.620	0.620	0.637
b_3	μ_1	0.014	0.225	0.352	0.495	0.604	0.697	1.097	1.548	1.895	2.075
	$\sqrt{\mu_2}$	1.027	1.027	1.027	1.027	1.027	1.027	1.027	1.027	1.027	1.027
	μ_3	0.275	0.275	0.275	0.275	0.275	0.275	0.275	0.275	0.275	0.275
	μ_4	0.091	0.091	0.091	0.091	0.091	0.091	0.091	0.091	0.091	0.091
δ_3	μ_1	0.007	0.079	0.138	0.222	0.299	0.372	0.748	1.230	1.608	1.804
	$\sqrt{\mu_2}$	0.540	0.559	0.586	0.624	0.657	0.685	0.798	0.881	0.919	0.934
	μ_3	0.527	0.761	0.859	0.928	0.949	0.942	0.727	0.516	0.427	0.396
	μ_4	2.294	2.437	2.574	2.363	1.963	1.596	0.615	0.274	0.185	0.159
b_4	μ_1	0.019	0.230	0.356	0.499	0.609	0.701	1.102	1.553	1.899	2.079
	$\sqrt{\mu_2}$	0.969	0.969	0.969	0.969	0.969	0.969	0.969	0.969	0.969	0.969
	μ_3	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129
	μ_4	0.056	0.056	0.056	0.056	0.056	0.056	0.056	0.056	0.056	0.056
δ_4	μ_1	-0.016	0.056	0.115	0.200	0.278	0.351	0.733	1.223	1.605	1.803
	$\sqrt{\mu_2}$	0.524	0.540	0.564	0.595	0.621	0.644	0.744	0.821	0.859	0.874
	μ_3	0.025	0.469	0.653	0.730	0.757	0.763	0.599	0.392	0.294	0.261
	μ_4	3.222	3.003	2.745	2.328	1.928	1.601	0.628	0.272	0.177	0.149

Table 3.1
Summary Statistics
Monte Carlo

	R^2	.00001	.010	.025	.050	.075	.100	.250	.500	.750	.900
β_1		.0069	.218	.345	.487	.596	.689	1.089	1.541	1.887	2.067
b_5	μ_1	0.061	0.272	0.399	0.542	0.651	0.744	1.144	1.595	1.942	2.122
	$\sqrt{\mu_2}$	1.015	1.015	1.015	1.015	1.015	1.015	1.015	1.015	1.015	1.015
	μ_3	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007
	μ_4	0.245	0.245	0.245	0.245	0.245	0.245	0.245	0.245	0.245	0.245
δ_5	μ_1	0.028	0.105	0.165	0.250	0.327	0.400	0.778	1.265	1.647	1.844
	$\sqrt{\mu_2}$	0.535	0.548	0.566	0.597	0.628	0.657	0.772	0.858	0.900	0.915
	μ_3	0.091	0.491	0.655	0.761	0.763	0.725	0.478	0.263	0.165	0.131
	μ_4	3.116	3.051	2.886	2.531	2.134	1.787	0.733	0.358	0.285	0.269
b_6	μ_1	0.008	0.219	0.346	0.488	0.598	0.690	1.091	1.542	1.888	2.069
	$\sqrt{\mu_2}$	0.981	0.981	0.981	0.981	0.981	0.981	0.981	0.981	0.981	0.981
	μ_3	-0.105	-0.105	-0.105	-0.105	-0.105	-0.105	-0.105	-0.105	-0.105	-0.105
	μ_4	-0.249	-0.249	-0.249	-0.249	-0.249	-0.249	-0.249	-0.249	-0.249	-0.249
δ_6	μ_1	0.004	0.082	0.141	0.226	0.304	0.377	0.751	1.229	1.605	1.801
	$\sqrt{\mu_2}$	0.504	0.499	0.520	0.557	0.589	0.617	0.733	0.827	0.871	0.887
	μ_3	-0.350	0.055	0.178	0.271	0.333	0.367	0.285	0.103	0.019	-0.008
	μ_4	2.122	1.653	1.742	1.360	0.906	0.560	-0.203	-0.361	-0.352	-0.340
b_7	μ_1	0.031	0.242	0.369	0.511	0.621	0.713	1.114	1.565	1.912	2.092
	$\sqrt{\mu_2}$	0.976	0.976	0.976	0.976	0.976	0.976	0.976	0.976	0.976	0.976
	μ_3	0.046	0.046	0.046	0.046	0.046	0.046	0.046	0.046	0.046	0.046
	μ_4	0.418	0.418	0.418	0.418	0.418	0.418	0.418	0.418	0.418	0.418
δ_7	μ_1	0.010	0.093	0.152	0.234	0.309	0.380	0.755	1.239	1.619	1.816
	$\sqrt{\mu_2}$	0.577	0.562	0.559	0.575	0.598	0.621	0.730	0.817	0.859	0.875
	μ_3	-0.150	0.502	0.584	0.612	0.622	0.614	0.436	0.224	0.132	0.103
	μ_4	5.704	4.500	3.393	2.714	2.296	1.936	0.760	0.368	0.307	0.298
b_8	μ_1	0.088	0.299	0.425	0.568	0.677	0.770	1.170	1.622	1.968	2.148
	$\sqrt{\mu_2}$	1.035	1.035	1.035	1.035	1.035	1.035	1.035	1.035	1.035	1.035
	μ_3	0.054	0.054	0.054	0.054	0.054	0.054	0.054	0.054	0.054	0.054
	μ_4	-0.018	-0.018	-0.018	-0.018	-0.018	-0.018	-0.018	-0.018	-0.018	-0.018
δ_8	μ_1	0.048	0.130	0.192	0.278	0.356	0.429	0.808	1.295	1.675	1.872
	$\sqrt{\mu_2}$	0.554	0.571	0.595	0.628	0.659	0.687	0.802	0.888	0.929	0.944
	μ_3	0.071	0.509	0.591	0.621	0.644	0.642	0.480	0.290	0.198	0.166
	μ_4	2.472	2.697	2.569	2.094	1.643	1.280	0.320	0.014	-0.040	-0.050

Table 5.2
Averages of the Summary Statistics
Monte Carlo

R^2	.00001	.010	.025	.050	.075	0.10	0.25	0.50	0.75	0.90
Least Squares										
$E(b)$	0.006	0.218	0.345	0.487	0.596	0.689	1.089	1.541	1.887	2.067
$\bar{\mu}_1$	0.023	0.235	0.361	0.504	0.615	0.706	1.106	1.558	1.904	2.083
$E(\hat{\sigma})$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$\sqrt{\mu}_2$	1.003	1.003	1.003	1.003	1.003	1.003	1.003	1.003	1.003	1.003
James-Stein										
$E(\delta)$	0.002	0.073	0.128	0.207	0.280	0.351	0.724	1.209	1.590	1.787
$\bar{\mu}_1$	0.004	0.082	0.142	0.226	0.304	0.376	0.754	1.237	1.616	1.813
$E(\hat{\sigma}_\delta)$	0.559	0.568	0.583	0.609	0.633	0.657	0.760	0.845	0.886	0.902
$\sqrt{\mu}_2$	0.538	0.548	0.566	0.598	0.626	0.652	0.760	0.846	0.887	0.903
Least Squares										
$\bar{\mu}_1 - E(b)$	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017
$\sqrt{\mu}_2 / E(\hat{\sigma})$	1.003	1.003	1.003	1.003	1.003	1.003	1.003	1.003	1.003	1.003
James-Stein										
$\bar{\mu}_1 - E(\delta)$	0.002	0.009	0.014	0.019	0.024	0.025	0.030	0.028	0.026	0.026
$\sqrt{\mu}_2 / E(\hat{\sigma}_\delta)$	0.962	0.964	0.972	0.981	0.988	0.990	1.000	1.001	1.001	1.001

HISTOGRAM OF A TYPICAL LEAST SQUARES ESTIMATE
FROM THE MONTE CARLO

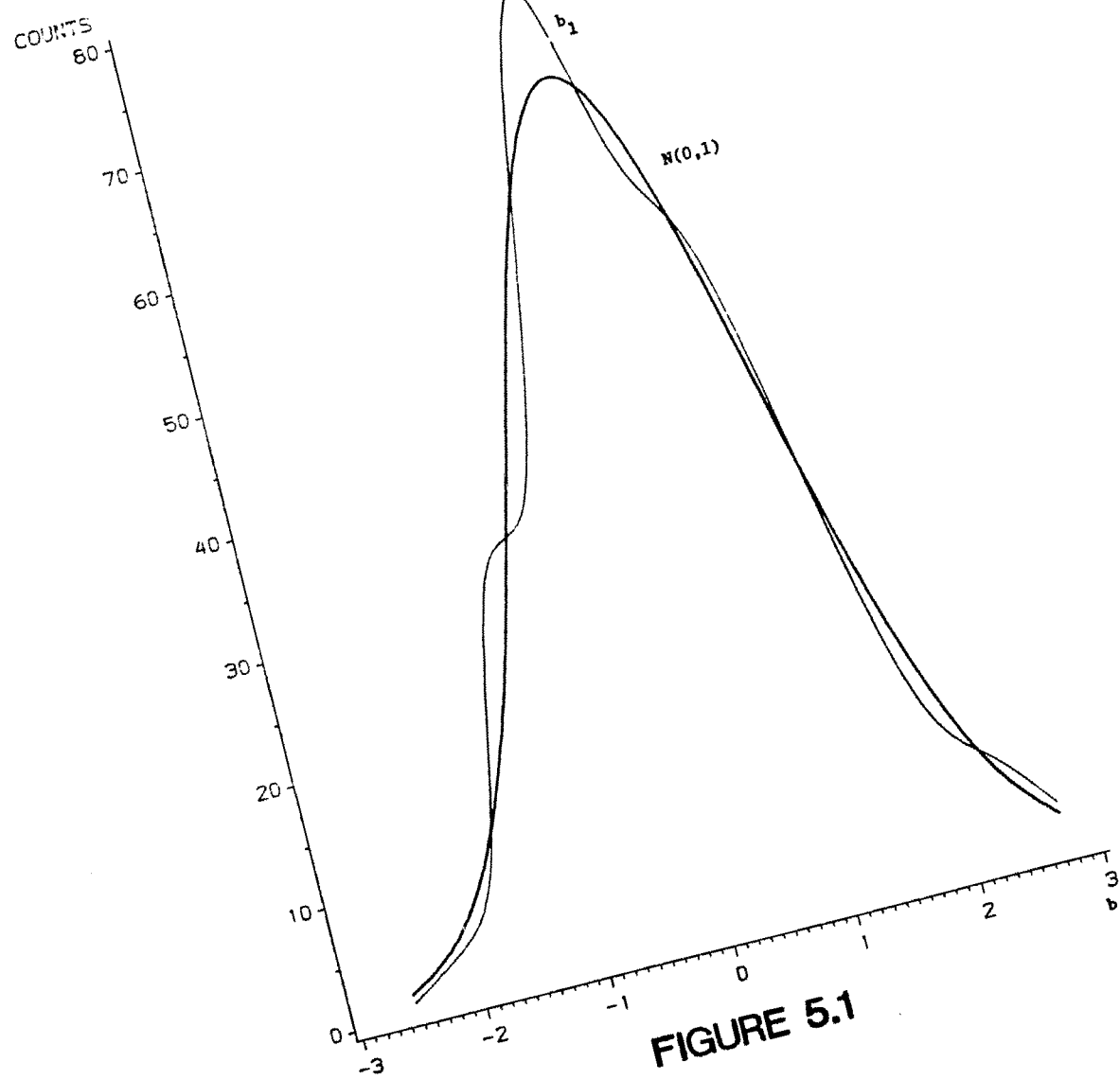


FIGURE 5.1

R-SQUARE = .00001

HISTOGRAM OF A TYPICAL JAMES-STEIN ESTIMATE
FROM THE MONTE CARLO

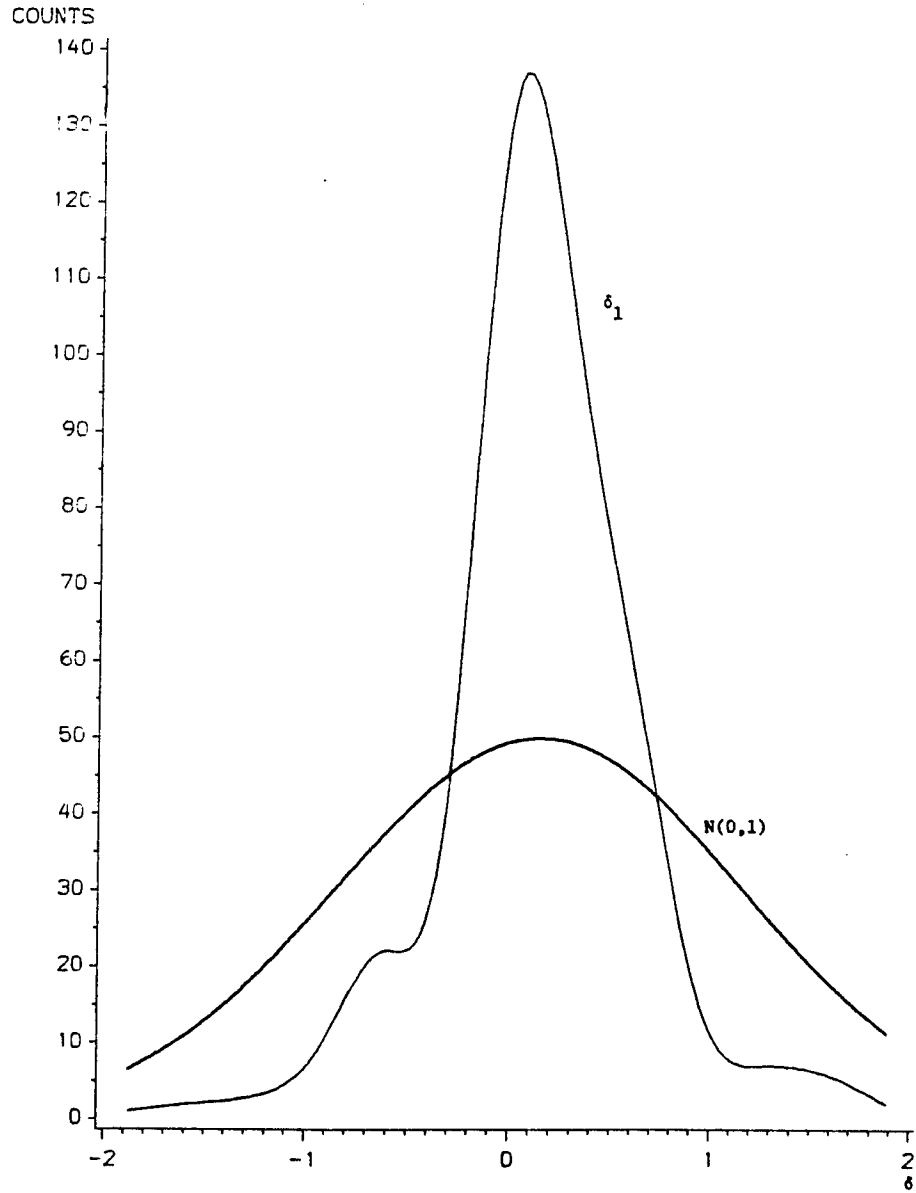


FIGURE 5.2

R-SQUARE= .00001

LEAST SQUARES AND JAMES-STEIN HISTOGRAMS
FROM THE MONTE CARLO

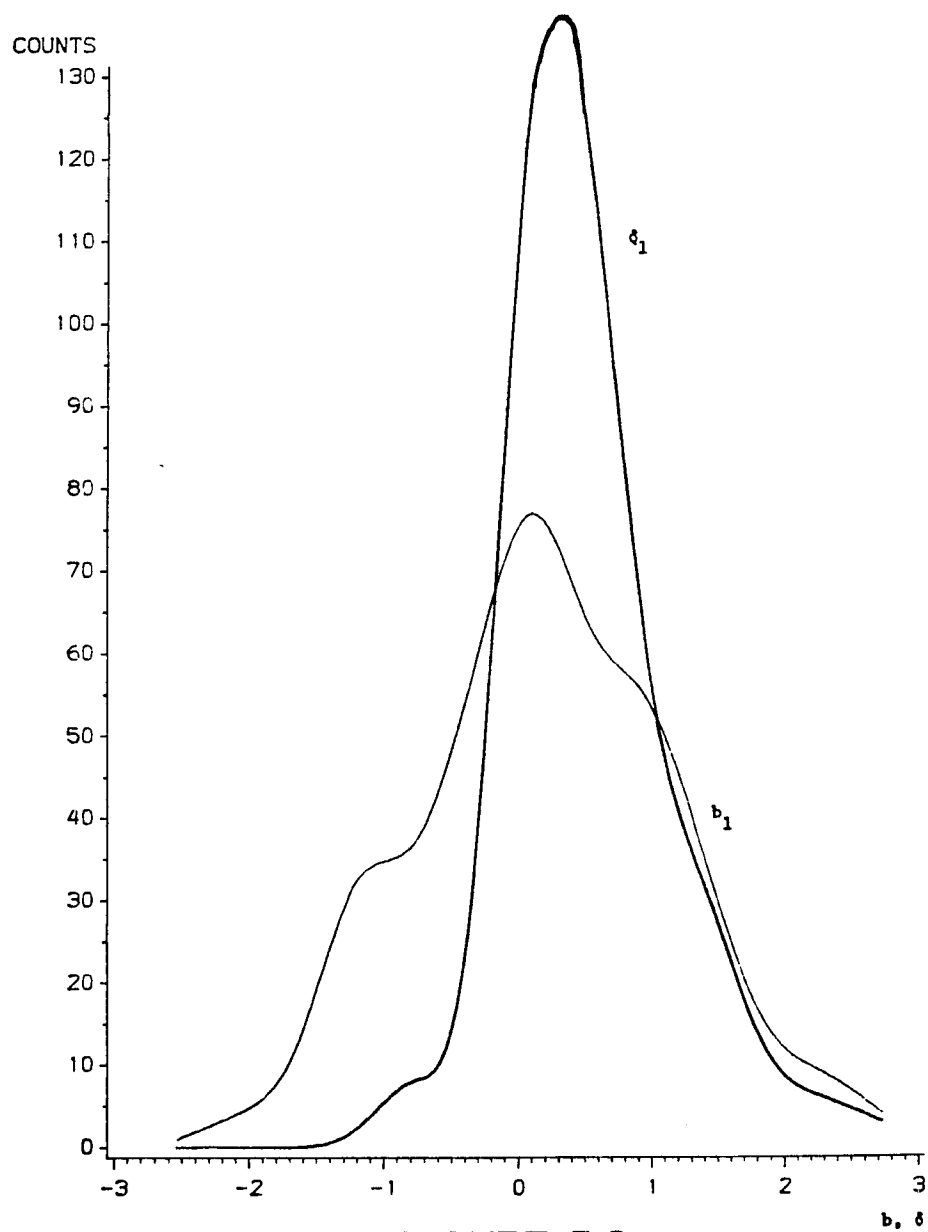


FIGURE 5.3

R-SQUARE=.00001

HISTOGRAMS OF TYPICAL JAMES-STEIN ESTIMATES
FROM THE MONTE CARLO

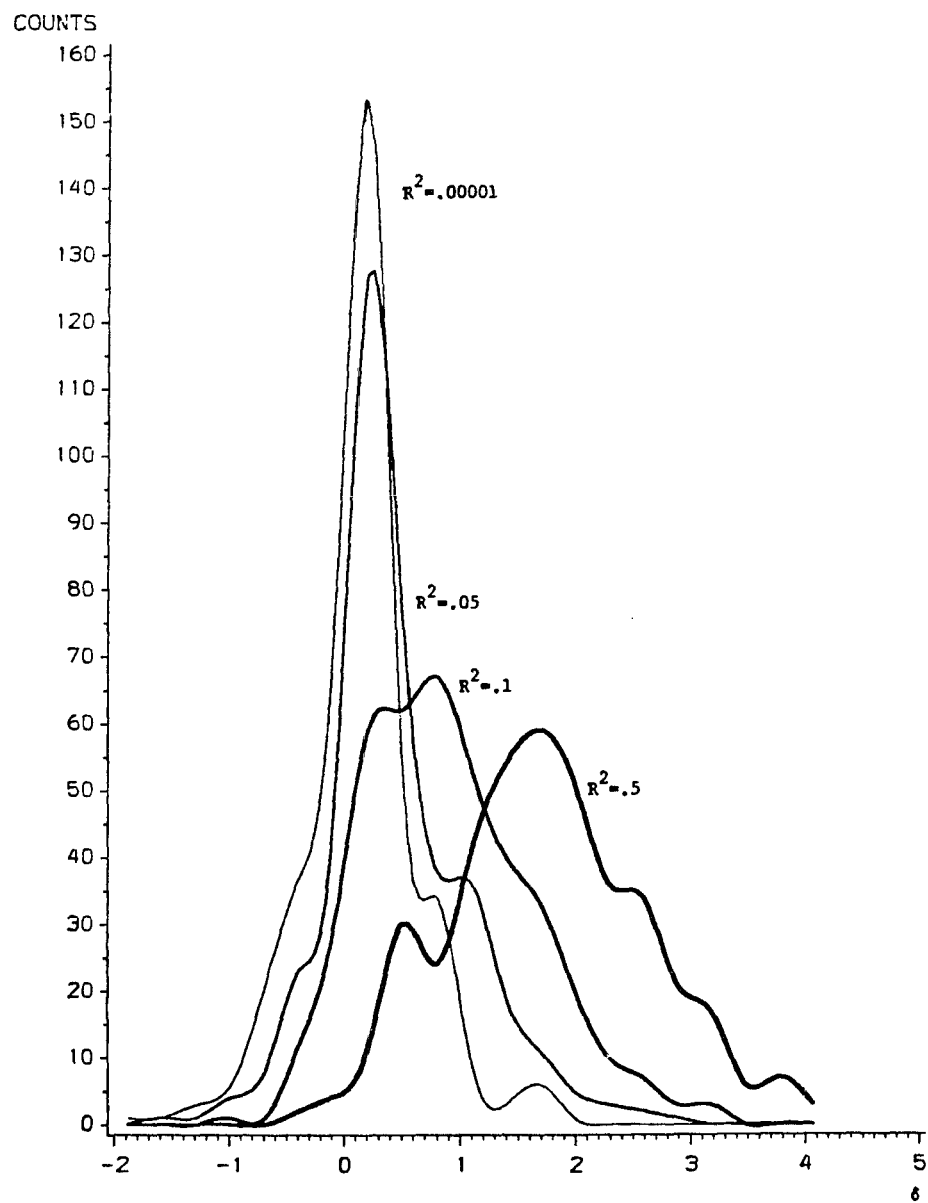


FIGURE 5.4

R-SQUARE=.00001, .05, .1, AND .5

The 5% and 10% critical values (Stephens, 1974) are .895 and .819, respectively. The hypothesis that b_i are normally distributed cannot be rejected based on this test.

Table 5.3
Kolmogorov's D Statistics
Monte Carlo

Estimate	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8
D	.021	.030	.035	.028	.034	.024	.026	.025

In the next section, Efron's (1979) bootstrap is discussed and bootstrap estimates of b , δ , and their standard errors are obtained. Then, percentile intervals are derived for least squares and James-Stein estimators.

5.4 The Bootstrap

In its most common form, the bootstrap is a computer intensive technique which permits one to assess the variability of an estimate using the sample. Efron (1979, 1982), Freedman (1981), Bickel and Freedman (1981, 1983) and Freedman and Peters (1984) provide an overview of the technique and its properties.

Freedman and Peters (1984) describe the basic idea behind the technique in the following way:

In brief, the model has been fitted to data by some statistical procedure; and there are residuals, namely the difference between observed and fitted values. Some stochastic structure was imposed on the stochastic disturbance terms, explicitly or implicitly, in the fitting. The key idea is to resample the residuals, preserving this stochastic structure, so the standard errors are generated using the model's own assumptions.

Thus, let $\{y_1, \dots, y_T\}$ is a random sample of size T from an unknown probability density F , then let $T(y)$ be a statistic

distributed on F . Since F is unknown, it is estimated using the empirical c.d.f. \hat{F} which is obtained by assigning a mass of $1/T$ to each observation in y . Then, a bootstrap sample y^* is obtained by independently drawing, with replacement, T observations from \hat{F} . The statistic of interest $T(y^*)$ is recomputed using the bootstrap sample y^* . A large number N of bootstrap samples are taken and $T(y^*)$ is computed for each. The approximate c.d.f. of $T(y)$ is derived by choosing some critical point $0 < c < N$ and calculating the number (#) of the $i=1, \dots, N$ bootstrap statistics $T_i(y^*)$ that are less than or equal to the critical value c and dividing this quantity by the total number of bootstrap samples used (N), i.e.,

$$\hat{F}(c) = \#\{T_i(y^*) \leq c\} / N \quad i=1, \dots, N.$$

In effect, the model is being refitted to the "pseudo-data" generated by resampling and the statistic of interest is calculated for each pseudo-sample. The empirical distribution of this statistic is then being used to approximate its sampling distribution.

The bootstrap method described above can be used to generate approximate $100(1-\alpha)\%$ confidence intervals in the regression model and is commonly referred to as Efron's (1979, 1982) percentile method. In order to use the percentile bootstrap to obtain approximate confidence intervals centered at the OLS estimator, one first estimates the parameters of the model (5.3.1) using least squares and the available data. The least squares

estimates b are used to generate the set of residuals $\hat{e} = y - Xb$ which serve as estimates of the true disturbances of the linear model. These are believed to capture the underlying stochastic structure of the model. Note that the length of the least squares residual will be shorter than that of the true disturbance, i.e., $|\hat{e}_t| \leq |e_t|$. This problem may be adjusted for in two ways.

Wu (1985) has suggested that the least squares residuals be inflated by the factor $[T/(T-K)]^{\frac{1}{2}}$ (i.e., $\hat{e}_t^* = \hat{e}_t [T/(T-K)]^{\frac{1}{2}}$). This rescaling follows from the fact that

$$\begin{aligned} \text{trace}\{E[\hat{e}\hat{e}']\} &= \text{trace}\{E[e(I-X(X'X)^{-1}X')e']\} \\ &= \sigma^2(T-K) \end{aligned}$$

while

$$\text{trace}\{E[ee']\} = \sigma^2 T.$$

after rescaling however,

$$\text{trace}\{E[\hat{e}^*\hat{e}^{*'}]\} = \sigma^2 T = \text{trace}\{E[ee']\}.$$

On average this rescaling assures that we will obtain an accurate measure of the desired covariance matrix.

Another possibility is to studentize the residuals. That is, we can use the fact that $\hat{e}\hat{e}' = e(I-X(X'X)^{-1}X')e'$ which has t^{th} diagonal element $e_t(1-x_t(X'X)^{-1}x_t')e_t$. An appropriate scaling can be obtained by using

$$\hat{e}_t^* = (1-x_t(X'X)^{-1}x_t')^{-\frac{1}{2}} \hat{e}_t.$$

For i.i.d. errors, the two methods will be equivalent on average.

An i.i.d. sample $\hat{e}^* = \{\hat{e}_t^*\}_1^T$ is then drawn from the scaled residuals \hat{e}_t^* . The bootstrap observation $y_t^* = x_t'b + \hat{e}_t^*$

is obtained by treating b as the true population parameters and \hat{e}^* as the population errors. At this point, a single bootstrap least squares estimate is computed using

$$b^* = (X'X)^{-1}X'y^* \quad (5.4.1)$$

where the $T \times 1$ vector of bootstrap observations is denoted $y^* = Xb + \hat{e}^*$. Then, a large number N of random bootstrap samples $\{\hat{e}_t^*\}_1^T$ are drawn from which $\{y^*\}_1^N$ and $\{b^*\}_1^N$ are computed. Similarly, the bootstrap estimates of the James-Stein estimator is computed using

$$\delta^*(b^*) = [1 - cK\hat{\sigma}^{*2}/b^{*'}X'Xb^*]b^* \quad (5.4.2)$$

where $\hat{\sigma}^{*2} = (y^* - Xb^*)'(y^* - Xb^*)/(T-K)$. This yields the sequence $\{\delta^*\}_1^N$.

5.4.1 Bootstrap Estimates

For each of the 400 Monte Carlo estimates b , a set of bootstrap statistics can be computed. We have chosen to studentize the LS residuals using $[1 - x_i(X'X)^{-1}x_i']^{-\frac{1}{2}}$ as described in the preceding section. A bootstrap sample of size $T=30$ is drawn from the studentized LS residuals using SAS (1986, Version 5.16) RANUNI. For a single bootstrap sample draw 30 uniform random deviates on the interval $[0,1]$, multiply each by 30, and round the resulting number up to the nearest integer. This yields 30 random integers on the interval $[1,30]$. If the integer 1 appears twice, then the 1st studentized residual is used twice in the bootstrap sample. Then use $\{\hat{e}_t^*\}_1^T$ to compute the bootstrap sample $\{y^*\}_1^T$ and to generate b^* and δ^* . Using this procedure, $N=500$ bootstrap samples are drawn for each of

M=400 Monte Carlo samples. The entire process is repeated for each parameter point implied by the use of (5.2.2).

The sample characteristics of the estimators b^* and δ^* are examined below. In Table 5.4, we report the mean, standard deviation, skewness and kurtosis of each element of b^* and δ^* using

$$\kappa_{m,i} = N^{-1} \sum_{n=1}^N \tilde{b}_{n,i}$$

$$\mu_{1i} = M^{-1} \sum_{m=1}^M \kappa_{m,i}$$

$$\mu_{2i} = (NM)^{-1} \sum_{n=1}^N \sum_{m=1}^M (\tilde{b}_{nm,i} - \kappa_{m,i})^2$$

$$\mu_{3i} = (NM)^{-1} \sum_{n=1}^N \sum_{m=1}^M (\tilde{b}_{nm,i} - \kappa_{m,i})^3 / \left[\sum_{n=1}^N \sum_{m=1}^M (\tilde{b}_{nm,i} - \kappa_{m,i})^2 \right]^{3/2}$$

$$\mu_{4i} = \{ (NM)^{-1} \sum_{n=1}^N \sum_{m=1}^M (\tilde{b}_{nm,i} - \kappa_{m,i})^4 / \left[\sum_{n=1}^N \sum_{m=1}^M (\tilde{b}_{nm,i} - \kappa_{m,i})^2 \right]^2 \} - 3$$

where $\tilde{b}_{n,i}$ is the estimate of β_i obtained from the n^{th} iteration of the bootstrap for a given Monte Carlo estimate b , and $\tilde{b}_{nm,i}$ is the estimate of β_i from the m^{th} Monte Carlo and the n^{th} iteration of the bootstrap. Table 5.4 gives us some idea of how the empirical distributions of b^* and δ^* compare to one another.

In Table 5.4 we see that the bootstrap averages are much more stable than they are over the Monte Carlo (i.e., Table 5.1). For least squares the bootstrap averages are nearly unbiased, have slightly smaller standard errors than expected (by about 2.5%), and have negligible degrees of

skewness and kurtosis. There appears to be little deviation from normality.

The average for each of the James-Stein estimates δ_i for $R^2 = .00001$ is also approximately equal to its expected value, and hence unbiased. Note however that standard errors are on average considerably overestimated by the bootstrap for small values of R^2 . The sample skewness of δ_i^* increases as R^2 increases, reaching a maximum at $R^2 = .25$. From there, it declines and becomes negligible at $R^2 = .90$. Kurtosis at the origin is approximately equal to one, which is significantly lower than the kurtosis reported for the Monte Carlo values of δ . Kurtosis declines uniformly to 0 as $R^2 \rightarrow 1$.

In Table 5.5 we average the results of Table 5.4 across elements (i.e., $\bar{\mu}_p = \sum_{i=1}^K \mu_{pi} / K$ $p=1,2$) and report the difference between b and its expectation as well as the ratio of estimated to expected standard error. This yields overall information about how well the bootstrap approximates the expectations of b and δ and their standard errors in repeated sampling. Thus, the evaluation of μ_1 and μ_2 across Monte Carlo, bootstrap, and elements is a concise way to summarize the accuracy of the bootstrap.

Note from the bottom portion of Table 5.5 that in the bootstrap least squares is approximately unbiased (.017) and that the standard error of b is underestimated by about 2.5%. From Table 5.5 it is also apparent that the in bootstrap μ_1 and $\sqrt{\mu_2}$ associated with the James-Stein rule

do not always conform to their expected values. Although estimates at the origin are unbiased, the Stein estimates from the bootstrap exceed their expected values by as much as .03.; the degree of overestimation increases as $R^2 \rightarrow .25$ and declines thereafter. For standard errors, the bootstrap average at $R^2 = .00001$ is 0.707 while the expected value is 0.559. Hence, at the origin, the variability of the bootstrap Stein-rule estimates is on average 26% greater than it should be. This degree of overestimation declines, however, as $R^2 \rightarrow .5$; thereafter, standard error is actually overestimated (1.5% at $R^2 = .9$).

This apparent failure of the bootstrap to accurately capture the variability of the James-Stein estimator is puzzling, especially since it does appear to capture the distributions of the James-Stein estimator's components, b and u .

For $R^2 = .00001$, smoothed histograms for typical sequences $\{b_i^*\}$ and $\{\delta_i^*\}$ appear in Figure 5.5. Although these show the same general characteristics as the Monte Carlo histograms shown in Figure 5.2, the density of δ_i^* near 0 is smaller than that of δ_i and the histogram of δ_i^* has fatter tails.

Table 5.4
Summary Statistics
Bootstrap

	R^2	.00001	.010	.025	.050	.075	.100	.250	.500	.750	.900
β_1		.0069	.218	.345	.487	.596	.689	1.089	1.541	1.887	2.067
b_1	μ_1	-0.008	0.202	0.328	0.471	0.581	0.673	1.073	1.525	1.871	2.051
	$\sqrt{\mu_2}$	0.967	0.967	0.967	0.967	0.967	0.967	0.967	0.967	0.967	0.967
	μ_3	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004
	μ_4	-0.024	-0.024	-0.024	-0.024	-0.024	-0.024	-0.024	-0.024	-0.024	-0.024
δ_1	μ_1	-0.003	0.136	0.223	0.327	0.411	0.484	0.835	1.276	1.631	1.818
	$\sqrt{\mu_2}$	0.698	0.704	0.712	0.724	0.735	0.745	0.792	0.840	0.868	0.880
	μ_3	-0.002	0.063	0.102	0.137	0.154	0.165	0.180	0.150	0.117	0.101
	μ_4	0.995	0.897	0.837	0.756	0.664	0.576	0.289	0.109	0.041	0.021
b_2	μ_1	-0.021	0.189	0.316	0.458	0.568	0.660	1.061	1.512	1.859	2.039
	$\sqrt{\mu_2}$	0.980	0.980	0.980	0.980	0.980	0.980	0.980	0.980	0.980	0.980
	μ_3	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004
	μ_4	-0.029	-0.029	-0.029	-0.029	-0.029	-0.029	-0.029	-0.029	-0.029	-0.029
δ_2	μ_1	-0.025	0.116	0.205	0.310	0.395	0.470	0.823	1.265	1.621	1.808
	$\sqrt{\mu_2}$	0.712	0.719	0.727	0.739	0.750	0.760	0.807	0.854	0.882	0.894
	μ_3	-0.011	0.055	0.095	0.130	0.149	0.162	0.174	0.139	0.107	0.092
	μ_4	0.975	0.873	0.803	0.687	0.584	0.514	0.252	0.091	0.026	0.008
b_3	μ_1	0.012	0.223	0.349	0.492	0.602	0.694	1.094	1.546	1.892	2.072
	$\sqrt{\mu_2}$	0.979	0.979	0.979	0.979	0.979	0.979	0.979	0.979	0.979	0.979
	μ_3	-0.006	-0.006	-0.006	-0.006	-0.006	-0.006	-0.006	-0.006	-0.006	-0.006
	μ_4	-0.036	-0.036	-0.036	-0.036	-0.036	-0.036	-0.036	-0.036	-0.036	-0.036
δ_3	μ_1	0.011	0.153	0.241	0.345	0.429	0.503	0.855	1.297	1.652	1.839
	$\sqrt{\mu_2}$	0.711	0.715	0.722	0.733	0.744	0.754	0.802	0.851	0.880	0.892
	μ_3	-0.020	0.042	0.085	0.124	0.147	0.163	0.180	0.145	0.110	0.093
	μ_4	0.969	0.882	0.811	0.701	0.620	0.556	0.276	0.094	0.022	0.002
b_4	μ_1	0.023	0.234	0.361	0.504	0.613	0.706	1.106	1.557	1.904	2.084
	$\sqrt{\mu_2}$	0.978	0.978	0.978	0.978	0.978	0.978	0.978	0.978	0.978	0.978
	μ_3	-0.006	-0.006	-0.006	-0.006	-0.006	-0.006	-0.006	-0.006	-0.006	-0.006
	μ_4	-0.025	-0.025	-0.025	-0.025	-0.025	-0.025	-0.025	-0.025	-0.025	-0.025
δ_4	μ_1	0.013	0.152	0.239	0.342	0.426	0.500	0.853	1.298	1.656	1.844
	$\sqrt{\mu_2}$	0.707	0.712	0.720	0.733	0.745	0.755	0.804	0.852	0.880	0.892
	μ_3	0.010	0.072	0.111	0.143	0.160	0.172	0.186	0.147	0.111	0.094
	μ_4	1.059	0.897	0.806	0.727	0.659	0.577	0.272	0.092	0.031	0.013

Table 3.4
Summary Statistics
Bootstrap

	R^2	.00001	.010	.025	.050	.075	.100	.250	.500	.750	.900
β_i		.0069	.218	.345	.487	.596	.689	1.089	1.541	1.887	2.067
b_5	μ_1	0.063	0.274	0.401	0.544	0.653	0.746	1.146	1.598	1.944	2.124
	$\sqrt{\mu_2}$	0.973	0.973	0.973	0.973	0.973	0.973	0.973	0.973	0.973	0.973
	μ_3	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009
	μ_4	-0.026	-0.026	-0.026	-0.026	-0.026	-0.026	-0.026	-0.026	-0.026	-0.026
δ_5	μ_1	0.043	0.183	0.272	0.376	0.461	0.535	0.891	1.337	1.695	1.883
	$\sqrt{\mu_2}$	0.705	0.712	0.720	0.733	0.744	0.754	0.803	0.851	0.878	0.890
	μ_3	0.029	0.092	0.126	0.157	0.175	0.186	0.203	0.163	0.126	0.109
	μ_4	0.980	0.869	0.807	0.746	0.671	0.595	0.322	0.097	0.032	0.014
b_6	μ_1	0.007	0.218	0.345	0.488	0.597	0.690	1.090	1.541	1.888	2.068
	$\sqrt{\mu_2}$	0.964	0.964	0.964	0.964	0.964	0.964	0.964	0.964	0.964	0.964
	μ_3	-0.002	-0.002	-0.002	-0.002	-0.002	-0.002	-0.002	-0.002	-0.002	-0.002
	μ_4	-0.023	-0.023	-0.023	-0.023	-0.023	-0.023	-0.023	-0.023	-0.023	-0.023
δ_6	μ_1	0.006	0.146	0.234	0.338	0.422	0.496	0.850	1.292	1.648	1.835
	$\sqrt{\mu_2}$	0.700	0.706	0.714	0.725	0.736	0.746	0.792	0.839	0.867	0.878
	μ_3	0.004	0.058	0.090	0.123	0.145	0.159	0.178	0.146	0.112	0.096
	μ_4	0.947	0.901	0.832	0.733	0.626	0.544	0.270	0.094	0.032	0.014
b_7	μ_1	0.027	0.238	0.364	0.507	0.617	0.709	1.109	1.561	1.907	2.087
	$\sqrt{\mu_2}$	0.970	0.970	0.970	0.970	0.970	0.970	0.970	0.970	0.970	0.970
	μ_3	-0.006	-0.006	-0.006	-0.006	-0.006	-0.006	-0.006	-0.006	-0.006	-0.006
	μ_4	-0.028	-0.028	-0.028	-0.028	-0.028	-0.028	-0.028	-0.028	-0.028	-0.028
δ_7	μ_1	0.015	0.155	0.243	0.347	0.431	0.505	0.859	1.303	1.660	1.848
	$\sqrt{\mu_2}$	0.701	0.708	0.716	0.728	0.740	0.750	0.798	0.846	0.873	0.885
	μ_3	0.015	0.089	0.125	0.154	0.170	0.180	0.187	0.145	0.110	0.094
	μ_4	0.928	0.863	0.798	0.700	0.622	0.548	0.290	0.094	0.031	0.012
b_8	μ_1	0.086	0.297	0.424	0.566	0.676	0.768	1.169	1.620	1.967	2.147
	$\sqrt{\mu_2}$	0.985	0.985	0.985	0.985	0.985	0.985	0.985	0.985	0.985	0.985
	μ_3	-0.002	-0.002	-0.002	-0.002	-0.002	-0.002	-0.002	-0.002	-0.002	-0.002
	μ_4	-0.008	-0.008	-0.008	-0.008	-0.008	-0.008	-0.008	-0.008	-0.008	-0.008
δ_8	μ_1	0.059	0.201	0.291	0.396	0.482	0.557	0.914	1.360	1.718	1.906
	$\sqrt{\mu_2}$	0.719	0.724	0.732	0.743	0.753	0.763	0.810	0.858	0.886	0.898
	μ_3	0.012	0.069	0.108	0.142	0.159	0.170	0.181	0.147	0.112	0.095
	μ_4	0.977	0.914	0.838	0.732	0.645	0.579	0.322	0.120	0.054	0.036

Table 5.5
Averages of the Summary Statistics
Bootstrap

R^2	.00001	.010	.025	.050	.075	0.10	0.25	0.50	0.75	0.90
Least Squares										
$E(b)$	0.006	0.218	0.345	0.487	0.596	0.689	1.089	1.541	1.887	2.067
$\bar{\mu}_1$	0.023	0.235	0.361	0.504	0.615	0.706	1.106	1.558	1.904	2.083
$E(\hat{\sigma})$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$\sqrt{\mu}_2$	0.975	0.975	0.975	0.975	0.975	0.975	0.975	0.975	0.975	0.975
James-Stein										
$E(\delta)$	0.002	0.073	0.128	0.207	0.280	0.351	0.724	1.209	1.590	1.787
$\bar{\mu}_1$	0.015	0.156	0.243	0.348	0.432	0.507	0.860	1.304	1.660	1.848
$E(\hat{\sigma}_\delta)$	0.559	0.568	0.583	0.609	0.633	0.657	0.760	0.854	0.886	0.902
$\sqrt{\mu}_2$	0.707	0.713	0.720	0.733	0.744	0.754	0.801	0.849	0.877	0.889
Least Squares										
$\bar{\mu}_1 - E(b)$	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017
$\sqrt{\mu}_2 / E(\hat{\sigma})$	0.975	0.975	0.975	0.975	0.975	0.975	0.975	0.975	0.975	0.975
James-Stein										
$\bar{\mu}_1 - E(\delta)$	0.013	0.083	0.115	0.141	0.152	0.156	0.136	0.095	0.070	0.053
$\sqrt{\mu}_2 / E(\hat{\sigma}_\delta)$	1.264	1.255	1.234	1.203	1.117	1.147	1.053	0.994	0.989	0.985

5.4.2 Percentile Confidence Intervals

Given the sequences of bootstrap statistics $\{b^*\}$ and $\{\delta^*\}$, approximate $100(1-\alpha)\%$ confidence intervals can now be computed using the percentile method.

In order to use the bootstrap to construct approximate confidence intervals for individual elements of β , the N values of the i^{th} element of $\{b^*\}_1^N$ are ranked in ascending order forming a bootstrap histogram. The $100(\alpha/2)\%$ and $100(1-\alpha/2)\%$ percentiles of the bootstrap histogram

$$b_i \in [b_i^*(\alpha/2), b_i^*(1-\alpha/2)] \quad i=1, \dots, K \quad (5.4.3)$$

are taken as an approximate $1-\alpha$ central confidence interval for the unknown parameter β_i .

A histogram of b_i^* from a single iteration of the Monte Carlo is shown in Figure 5.6. The 5% and 95% percentile points are given and demark the central 90% bootstrap confidence interval for β_i which is centered at b_i .

For $100(1-\alpha)\%$ confidence intervals centered at the James-Stein estimator the N values of i^{th} element of $\{\delta^*\}_1^N$ are ranked in ascending order forming a bootstrap histogram. The $100(\alpha/2)\%$ and $100(1-\alpha/2)\%$ percentiles of the bootstrap histogram

$$\delta_i \in [\delta_i^*(\alpha/2), \delta_i^*(1-\alpha/2)] \quad i=1, \dots, K \quad (5.4.4)$$

are taken as an approximate $1-\alpha$ central confidence interval for the unknown parameter β_i .

Similarly, a histogram for a single δ_i^* from one iteration of the Monte Carlo is shown in Figure 5.7. The 5% and 95% points are given and demark the central 90%

bootstrap confidence interval for β_i which are centered at δ_i .

The coverage probabilities and sizes of percentile bootstrap confidence intervals are reported in Tables 5.6 and 5.7 below.

As theory would predict, the size and coverage probability of confidence intervals and ellipsoids centered at least squares is independent of the location of the true parameters. This feature of LS confidence intervals is unaffected by use of the bootstrap procedure.

From Table 5.6 we can see that for least squares, 90% intervals derived using the percentile bootstrap (5.4.3) cover slightly less than 90% of the time and are slightly smaller than normal distribution theory would suggest. For instance the usual interval, $b \pm \sigma z_{\alpha/2}$, would have length equal to $2(z_{\alpha/2})\sigma$. For $\sigma=1$ and $\alpha=.1$, length would be 3.29. Thus percentile intervals are about 2.5% shorter than normal theory suggests and is due to the bootstrap's approximate 2.5% underestimation of the least squares estimator's standard error. Increasing the number of bootstrap samples from 500 to 1000 did not alleviate this problem.

Percentile bootstrap intervals centered at the James-Stein estimates performed reasonably well, especially near the origin. At $R^2=.00001$, these intervals cover 91%-94% of the time and are approximately 30% smaller than least squares percentile intervals. As we move away from the

origin, the intervals become larger and their coverage probability decreases. For $R^2 > .9$, coverage begins to increase as $\delta \rightarrow b$. Percentile intervals perform at their worst at $R^2 = .5$; on average they cover only 84%-88% of the time (whereas LS cover 87%-91%) and are about 15% shorter than LS. The relatively poor performance of biased confidence intervals centered at the James-Stein estimator is to be expected for intermediate values of R^2 . When R^2 is very small, estimates are nearly unbiased and have much lower variance. Hence, intervals will be smaller than ones based on LS and on average have the correct center. For very large values of R^2 , the estimates, standard errors, and confidence intervals converge to those associated with least squares. In essence, the value of the statistic u which controls the degree of shrinkage is unable to differentiate between good and bad prior information in the intermediate range of R^2 (i.e., for moderate degrees of noncentrality, the statistic has a relatively high probability of type II error). As a result, the center of the confidence interval is drawn away from the true value. The result is a confidence interval which is smaller than the least squares procedure, and which is centered on average at the wrong point.

From Table 5.7 we see that the 95% percentile bootstrap intervals for least squares have length 3.77-3.84 and cover 92% to 94% of the time. Again, theoretical intervals for this model have length 3.92, making the

bootstrap intervals about 2.5% too short.

The 95% intervals for the James-Stein estimates (Table 5.7) are similar to the 90% intervals reported in Table 5.6. At the 95% level, the percentile bootstrap (Table 5.7) appears to do a better job of getting intervals of the correct coverage probability near the origin. Seven of eight of the parameters are covered at least 96% of the time while being 25% shorter than percentile LS intervals. Furthermore, coverage of these intervals drops to about 92% around $R^2 = .25$ before increasing again. The 95% percentile bootstrap intervals centered at the James-Stein estimates apparently perform very well for the CNLRM.

Although it is difficult to make broad generalizations based on only two values of α , it appears that the bootstrap distribution of δ_1 is better at approximating its actual distribution further out in the tails p.d.f.

Other alternatives to the percentile intervals (5.4.3) and (5.4.4) can be obtained using the bootstrap. These intervals are implicitly based on bootstrap estimates of the covariance matrices $\text{Cov}(b)$ and $\text{Cov}(\delta)$. These are explored below. First, confidence ellipsoids are broadly defined and the bootstrap procedure is used to estimate these matrices. Bootstrap covariance estimates are then used to derive confidence intervals and ellipsoids centered at the least squares and the James-Stein estimators.

Table 5.6
Percentile Confidence Intervals
90%

R^2	.00001	.010	.025	.050	.075	.10	.25	.50	.75	.90
b_1 coverage	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88
b_1 length	3.18	3.18	3.18	3.18	3.18	3.18	3.18	3.18	3.18	3.18
δ_1 coverage	0.93	0.92	0.90	0.89	0.88	0.88	0.86	0.86	0.86	0.86
δ_1 length	2.28	2.30	2.32	2.36	2.40	2.43	2.59	2.75	2.85	2.89
b_2 coverage	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87
b_2 length	3.23	3.23	3.23	3.23	3.23	3.23	3.23	3.23	3.23	3.23
δ_2 coverage	0.91	0.91	0.88	0.86	0.85	0.85	0.84	0.84	0.84	0.84
δ_2 length	2.33	2.35	2.37	2.41	2.45	2.48	2.64	2.80	2.90	2.94
b_3 coverage	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
b_3 length	3.22	3.22	3.22	3.22	3.22	3.22	3.22	3.22	3.22	3.22
δ_3 coverage	0.91	0.91	0.89	0.87	0.86	0.85	0.85	0.84	0.84	0.84
δ_3 length	2.33	2.34	2.37	2.41	2.45	2.48	2.64	2.80	2.89	2.93
b_4 coverage	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91
b_4 length	3.21	3.21	3.21	3.21	3.21	3.21	3.21	3.21	3.21	3.21
δ_4 coverage	0.94	0.92	0.89	0.88	0.87	0.87	0.87	0.87	0.87	0.88
δ_4 length	2.30	2.32	2.35	2.39	2.43	2.47	2.63	2.79	2.88	2.92
b_5 coverage	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88
b_5 length	3.20	3.20	3.20	3.20	3.20	3.20	3.20	3.20	3.20	3.20
δ_5 coverage	0.93	0.93	0.91	0.91	0.90	0.89	0.88	0.88	0.88	0.88
δ_5 length	2.30	2.32	2.35	2.39	2.43	2.46	2.63	2.79	2.88	2.92
b_6 coverage	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86
b_6 length	3.17	3.17	3.17	3.17	3.17	3.17	3.17	3.17	3.17	3.17
δ_6 coverage	0.92	0.90	0.89	0.87	0.87	0.86	0.85	0.85	0.86	0.86
δ_6 length	2.28	2.30	2.32	2.36	2.40	2.43	2.59	2.75	2.84	2.88
b_7 coverage	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89
b_7 length	3.19	3.19	3.19	3.19	3.19	3.19	3.19	3.19	3.19	3.19
δ_7 coverage	0.94	0.92	0.91	0.90	0.89	0.89	0.87	0.88	0.88	0.88
δ_7 length	2.29	2.30	2.33	2.37	2.41	2.44	2.60	2.77	2.86	2.90
b_8 coverage	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87
b_8 length	3.23	3.23	3.23	3.23	3.23	3.23	3.23	3.23	3.23	3.23
δ_8 coverage	0.92	0.89	0.88	0.87	0.86	0.86	0.85	0.85	0.85	0.85
δ_8 length	2.34	2.36	2.38	2.42	2.45	2.49	2.65	2.81	2.90	2.94

Table 5.7
Percentile Confidence Intervals
95%

R^2		.00001	.010	.025	.050	.075	.10	.25	.50	.75	.90
b_1	coverage	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
	length	3.78	3.78	3.78	3.78	3.78	3.78	3.78	3.78	3.78	3.78
δ_1	coverage	0.96	0.96	0.96	0.94	0.94	0.93	0.93	0.93	0.93	0.93
	length	2.80	2.82	2.84	2.89	2.92	2.95	3.11	3.29	3.40	3.44
b_2	coverage	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92
	length	3.84	3.84	3.84	3.84	3.84	3.84	3.84	3.84	3.84	3.84
δ_2	coverage	0.96	0.95	0.94	0.94	0.93	0.93	0.92	0.92	0.93	0.93
	length	2.86	2.89	2.91	2.95	2.98	3.01	3.17	3.34	3.45	3.50
b_3	coverage	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92
	length	3.83	3.83	3.83	3.83	3.83	3.83	3.83	3.83	3.83	3.83
δ_3	coverage	0.96	0.95	0.94	0.94	0.93	0.93	0.92	0.92	0.93	0.93
	length	2.86	2.88	2.90	2.94	2.98	3.01	3.17	3.34	3.45	3.49
b_4	coverage	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
	length	3.84	3.84	3.84	3.84	3.84	3.84	3.84	3.84	3.84	3.84
δ_4	coverage	0.96	0.96	0.95	0.95	0.93	0.93	0.93	0.93	0.93	0.93
	length	2.84	2.86	2.88	2.92	2.96	2.99	3.15	3.33	3.44	3.49
b_5	coverage	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
	length	3.81	3.81	3.81	3.81	3.81	3.81	3.81	3.81	3.81	3.81
δ_5	coverage	0.96	0.96	0.95	0.95	0.94	0.94	0.93	0.94	0.94	0.94
	length	2.83	2.85	2.88	2.92	2.95	2.99	3.16	3.33	3.43	3.48
b_6	coverage	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92
	length	3.77	3.77	3.77	3.77	3.77	3.77	3.77	3.77	3.77	3.77
δ_6	coverage	0.96	0.95	0.94	0.93	0.93	0.92	0.90	0.91	0.91	0.91
	length	2.81	2.82	2.84	2.88	2.92	2.95	3.11	3.28	3.38	3.43
b_7	coverage	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
	length	3.80	3.80	3.80	3.80	3.80	3.80	3.80	3.80	3.80	3.80
δ_7	coverage	0.97	0.97	0.95	0.94	0.94	0.93	0.91	0.92	0.92	0.93
	length	2.82	2.84	2.87	2.91	2.94	2.97	3.14	3.32	3.42	3.47
b_8	coverage	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92
	length	3.86	3.86	3.86	3.86	3.86	3.86	3.86	3.86	3.86	3.86
δ_8	coverage	0.95	0.94	0.94	0.93	0.92	0.92	0.90	0.91	0.92	0.92
	length	2.88	2.90	2.92	2.96	2.99	3.03	3.19	3.37	3.47	3.52

LEAST SQUARES AND JAMES-STEIN HISTOGRAMS
FROM THE BOOTSTRAP

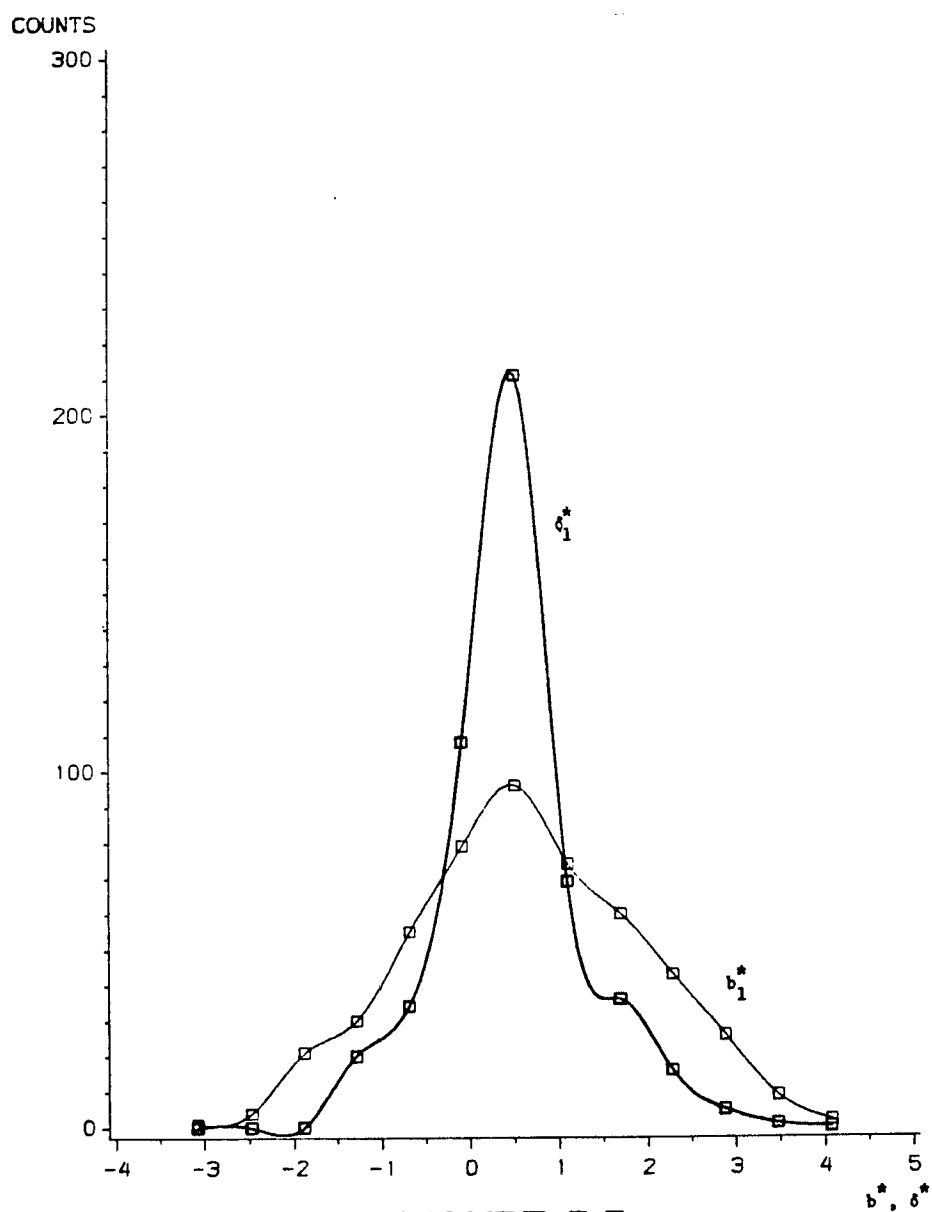


FIGURE 5.5

R-SQUARE=.00001

HISTOGRAM OF A TYPICAL LEAST SQUARES ESTIMATE
FROM THE BOOTSTRAP

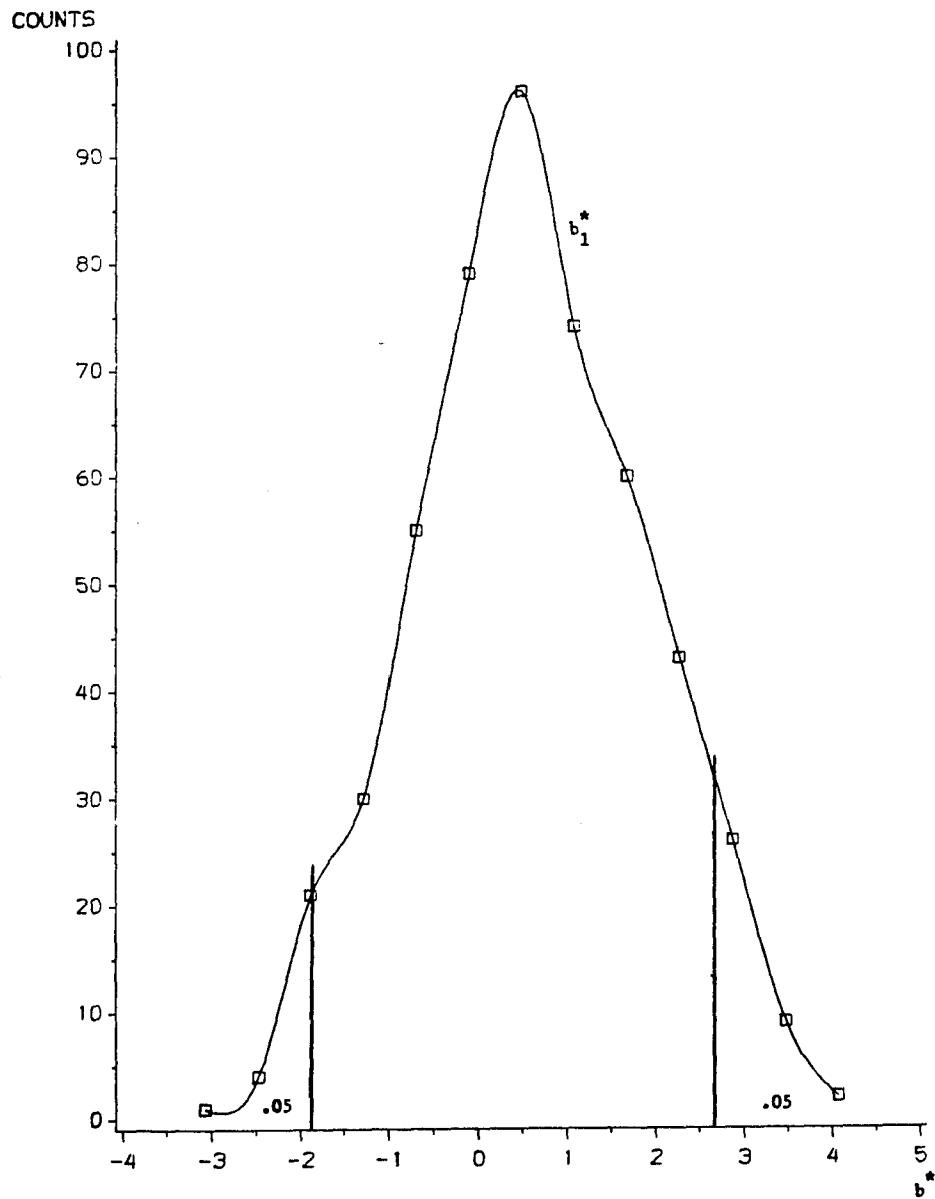


FIGURE 5.6

R-SQUARE= .00001

HISTOGRAM OF A TYPICAL JAMES-STEIN ESTIMATE
FROM THE BOOTSTRAP

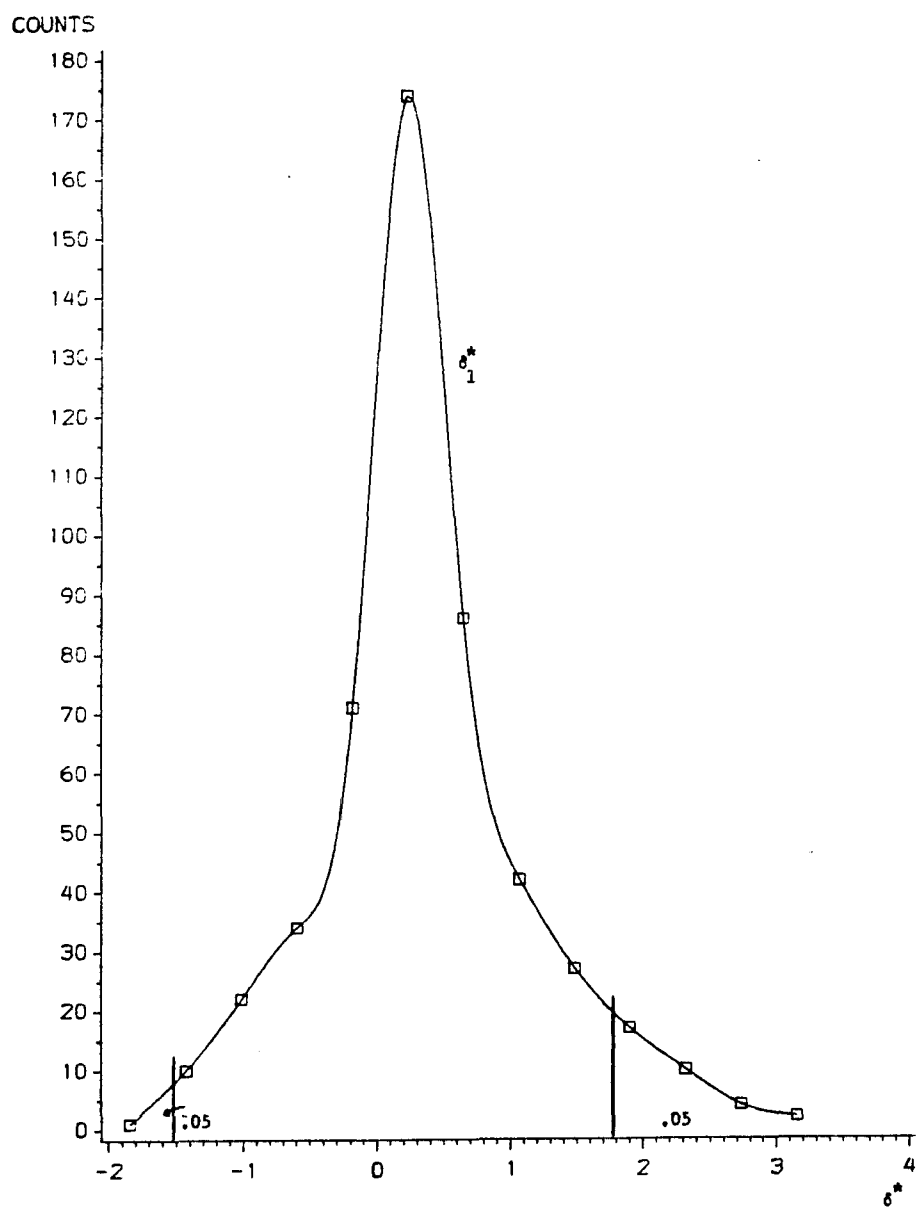


FIGURE 5.7

R-SQUARE= .00001

5.5 Confidence Ellipsoids

In this section confidence ellipsoids are broadly defined and confidence intervals derived as a special case. Then, three types of ellipsoids are suggested for use with LS and James-Stein estimators and an alternative to percentile intervals is discussed.

To begin, we define a random ellipsoid and discuss its size and the probability with which it covers the true values of the parameters. Let \tilde{b} again denote an arbitrary estimator of β , c an arbitrary positive constant, and A any positive definite weight matrix. The expression

$$Q = \{\beta : (\tilde{b} - \beta)' A^{-1} (\tilde{b} - \beta) \leq c\} \quad (5.5.1)$$

defines an ellipsoid centered at the estimator \tilde{b} . A $100(1-\alpha)\%$ confidence ellipsoid

$$C = \{c(1-\alpha) : \Pr[Q \leq c(1-\alpha)] \geq 1-\alpha\} \quad (5.5.2)$$

is obtained for $0 < \alpha < 1$ by finding the value of the constant $c(1-\alpha)$ which satisfies the inequality. This expression implies that the actual value of the parameter vector β will fall within the ellipsoid $100(1-\alpha)\%$ of the time in repeated samples.

In estimation of linear models like (5.3.1) the weight matrix A is usually chosen to be the covariance matrix of the estimator, i.e., $[\text{Cov}(\tilde{b})]$. This choice is made for several reasons. First, letting $A = [\text{Cov}(\tilde{b})]$ is reasonable in that $\text{Cov}(\tilde{b})$ measures the variability of the estimate about its expected value; in particular, diagonal elements measure the variability of individual \tilde{b}_i and off-diagonal

elements measure the linear association between \tilde{b}_i and \tilde{b}_j , $i \neq j$. Intuitively, if estimates are highly variable about their mean, then $[\text{Cov}(\tilde{b})]^{-1}$ will be small in the matrix sense and for given $c(1-\alpha)$, the set of β which satisfy (5.5.2) will be larger.

Another reason for using $A = [\text{Cov}(\tilde{b})]$ is that in many instances, the resulting quadratic form Q has a well-known probability distribution. For instance, if \tilde{b} is the MLE, then $\tilde{b} \sim N(\beta, \sigma^2 (X'X)^{-1})$, $(T-K)\hat{\sigma}^2 / \sigma^2 \sim \chi^2_{T-K}$ and \tilde{b} and $\hat{\sigma}^2$ are independent. The quadratic form $(\tilde{b} - \beta)' X'X (\tilde{b} - \beta) / K\hat{\sigma}^2 \sim F_{K, T-K}$, and c is chosen to be the $100(1-\alpha)\%$ point from this distribution.

Given $Q = [(\tilde{b} - \beta)' A^{-1} (\tilde{b} - \beta)]$, ellipsoids centered at linear combinations of the estimates \tilde{b} are easy to obtain. Such ellipsoids contain the set of all intervals for individual elements of β as a special case. If R is a $K \times 1$ vector of zeros with 1 in the i^{th} column, then $R(\tilde{b} - \beta) = (\tilde{b}_i - \beta_i)$. Intervals centered at \tilde{b}_i are formed using

$$q_i = (\tilde{b} - \beta)' R' [R A^{-1} R']^{-1} R (\tilde{b} - \beta)$$

or

$$q_i = (\tilde{b}_i - \beta_i)^2 a^{ii} \quad (5.5.3)$$

where a^{ii} is the i^{th} diagonal element of A^{-1} . For example, if \tilde{b}_i is the MLE of β_i and $\hat{\sigma}^2$ the m.v.u.e. of σ^2 , then $a^{ii} = (x_i' x_i) / \hat{\sigma}^2$ and (5.4.3) becomes

$$(\tilde{b}_i - \beta_i)^2 (x_i' x_i) / \hat{\sigma}^2 \sim t_{T-K}^2 \quad i=1, \dots, K$$

or

$$(\tilde{b}_i - \beta_i) (x_i' x_i)^{1/2} / \hat{\sigma} \sim t_{T-K}.$$

Thus,

$$b_i \pm \hat{\sigma}(x_i'x_i)^{-\frac{1}{2}} t_{\alpha/2} \quad (5.5.4)$$

is a $100(1-\alpha)\%$ confidence interval for β_i if $t_{\alpha/2}$ is the $\alpha/2$ point from the t_{T-K} distribution.

Smaller intervals and ellipsoids are to be preferred to larger ones, other things being equal. The volume of a K -dimensional ellipsoid in standard form can be measured as a function of the lengths of its semiaxes using

$$V_K = [\pi^{K/2}/\Gamma(K/2 + 1)] a_1 a_2 \dots a_K \quad (5.5.5)$$

where a_i is the length of the ellipsoid's i^{th} semiaxis. To obtain a measure of volume for the confidence ellipsoids derived below consider again the ellipsoid centered at \tilde{b}

$$\beta : (\tilde{b} - \beta)' A^{-1} (\tilde{b} - \beta) \leq c$$

and let $(\tilde{b} - \beta)' = s'$; thus, the transformation $s : s' A^{-1} s \leq c$ is used to center the ellipsoid at the origin. The ellipsoid can be further simplified by putting it into its canonical form. This is accomplished by defining the $K \times K$ matrix P such that $P'P = PP' = I_K$, $P'A^{-1}P = \Lambda$ and where $\Lambda = \text{diag}[\lambda_1, \dots, \lambda_K]$ is the diagonal matrix of characteristic roots of A^{-1} .

Putting $\theta' = s'P$ and $Z = P'A^{-\frac{1}{2}}$ enables us to express the ellipsoid in its canonical form

$$s'PP'A^{-\frac{1}{2}}A^{-\frac{1}{2}}PP's = \theta'\Lambda\theta \leq c$$

which implies

$$\theta_1^2(\lambda_1) + \theta_2^2(\lambda_2) + \dots + \theta_K^2(\lambda_K) \leq c.$$

The constant c is positive, hence the ellipsoid can be standardized without loss of generality by dividing both sides by c . This yields

$$\theta_1^2(\lambda_1/c) + \theta_2^2(\lambda_2/c) + \dots + \theta_K^2(\lambda_K/c) \leq 1.$$

Geometrically, $(\lambda_i/c)^{-1/2} = a_i$ represents the length of the i^{th} semiaxis of the ellipsoid. Hence, an ellipsoid's volume can be obtained using [Carter, Srivastava, Srivastava, and Ullah (1987)]

$$V_K = [n^{K/2}/\Gamma(K/2 + 1)] c^{K/2} \det(A)^{-1/2}$$

or

$$V_K = [n^{K/2}/\Gamma(K/2 + 1)] c^{K/2} \det(A^{-1/2}).$$

Thus,

$$V_K \propto c^{K/2} \det(A)^{-1/2}. \quad (5.5.6)$$

In the section below, we define several general forms of the ellipsoids to be considered. Then in section 5.6, the coverage probabilities and volumes of various estimates of these ellipsoids are explored.

5.5.1 Alternative Specifications of the Quadratic Form

Below, we consider three basic quadratic forms which are used to construct confidence ellipsoids for the unknown vector β .

The most commonly used quadratic form is based on the ellipsoid

$$Q_1 = (b - \beta)' [\text{Cov}(b)]^{-1} (b - \beta) / K. \quad (5.5.7)$$

Note that Q_1 is centered at the least squares estimator and its shape and volume are determined by $[\text{Cov}(b)]$, which must be estimated using the sample.

Two other quadratic forms are considered. These are

$$Q_2 = (\delta - \beta)' [\text{Cov}(b)]^{-1} (\delta - \beta) / K \quad (5.5.8)$$

and

$$Q_3 = (\delta - \hat{\beta})' [\text{Cov}(\delta)]^{-1} (\delta - \hat{\beta}) / K. \quad (5.5.9)$$

Although both of these ellipsoids are centered at the James-Stein estimator, the shape and volume of Q_2 is determined by the covariance of the least squares estimator while that of Q_3 is determined by the covariance of the James-Stein estimator. Because $\text{Cov}(b) - \text{Cov}(\delta) = \Delta$, a positive semi-definite matrix, ellipsoids centered at δ but measured in the metric $[\text{Cov}(b)]$ will generally be larger than those measured in $[\text{Cov}(\delta)]$.

The procedure described in the next section enables one to use bootstrap resampling to obtain an estimate of both the least squares and James-Stein covariance matrices. In section 5.5.3 a procedure for estimating $100(1-\alpha)\%$ confidence ellipsoids is presented. In addition, an alternative to the percentile intervals (5.4.3) and (5.4.4) is discussed which is based on bootstrap estimates of $\text{Cov}(b)$ and $\text{Cov}(\delta)$. Finally, in section 5.4.4 the procedure for obtaining approximate $100(1-\alpha)\%$ critical points for ellipsoids centered at LS and James-Stein estimators is presented.

5.5.2 Estimating Covariance

A natural extension of the bootstrap is to use it to estimate the variance-covariance matrix of an estimator, as Freedman and Peters (1984) have done for regression models with nonscalar covariance matrices. In this section the bootstrap is used to estimate $\text{Cov}(b)$ and $\text{Cov}(\delta)$ and then to obtain approximate $100(1-\alpha)\%$ critical points for the

resulting ellipsoids.

Under the assumptions of the model (5.3.1) and given \hat{e}^* represents a random sample from \hat{e} , $(b^* - b) \sim N(0, \hat{\sigma}^2(X'X)^{-1})$; therefore $B^* - \bar{B}^* \sim N(0, I_N \otimes \hat{\sigma}^2(X'X)^{-1})$ where B^* is the $N \times K$ matrix of bootstrap estimates whose columns are $\{b_i^*\}_1^N$, $\bar{B}^* = B^* - j_N \otimes \bar{b}^*$, j_N is a $N \times 1$ vector of ones, \otimes is the Kroneker product, \bar{b}^* is the $K \times 1$ vector whose elements are $\sum_{j=1}^N b_{ij}^* / N$, $i=1, \dots, K$, and N is the number of

bootstrap samples. Using this fact, the covariance matrix may be estimated using the bootstrap by computing

$$\text{Cov}(b^*) = (B^* - \bar{B}^*)' (B^* - \bar{B}^*) / (N-1) \sim W_K(N-1, \hat{\sigma}^2(X'X)^{-1}) \quad (5.5.10)$$

where W_K is the Wishart distribution with $N-1$ degrees of freedom and covariance $\hat{\sigma}^2(X'X)^{-1}$ [see Muirhead (1982), pp. 80-90 or Anderson (1984), Ch. 5].

Carter, Srivastava, Srivastava, and Ullah (1987) have devised an unbiased estimator of the mean square error (MSE) matrix of the James-Stein estimator. From this matrix, one can subtract the matrix $[\delta - E\delta][\delta - E\delta]'$ (i.e., $[\text{bias}(\delta)][\text{bias}(\delta)]'$) to obtain an estimate of covariance. Unfortunately their estimator can lead to negative definite estimates of the MSE and covariance matrices, especially when the degree of hypothesis error is relatively small. The CSSU estimator requires additional study and modification before it will be useful to practitioners.

The bootstrap can be used to estimate $\text{Cov}(\delta)$ in the

same way that it is used to estimate $\text{Cov}(b)$. Define

$$\text{Cov}(\delta^*) = (\Delta^* - \bar{\Delta}^*)' (\Delta^* - \bar{\Delta}^*) / (N-1) \quad (5.5.11)$$

where Δ^* is the $N \times K$ matrix of bootstrap estimates whose columns are $\{\delta_i^*\}_1^N$, $\bar{\Delta}^* = \Delta^* - j_N \otimes \bar{\delta}^*$, j_N is a $N \times 1$ vector of ones, \otimes is the Kroneker product, and $\bar{\delta}^*$ is the $K \times 1$ vector

whose elements are $\sum_{j=1}^N \delta_{ij}^* / N$, $i=1, \dots, K$. Like $\text{Cov}(b^*)$,

$\text{Cov}(\delta^*)$ is the empirically generated covariance obtained using the N values of the bootstrap. The sampling distribution of $[\text{Cov}(\delta^*)]$ is unknown; however, the approach will yield positive semidefinite estimates of the James-Stein estimator's covariance matrix.

The use of the bootstrap estimators $\text{Cov}(b^*)$ and $\text{Cov}(\delta^*)$ in the Monte Carlo study also provides an opportunity to assess their accuracy. We use the percent root-mean-square-error (PRMSE) measure to compare the diagonal elements of each estimated covariance matrix with their actual values. That is, let $[\hat{\text{Cov}}(\tilde{b})]$ represent the estimated value of $[\text{Cov}(\tilde{b})]$ and define PRMSE of $[\hat{\text{Cov}}(\tilde{b})]$ to be

$$\text{PRMSE } \hat{\text{Cov}}(\tilde{b}) = \left\{ K^{-1} \sum_{i=1}^K \frac{[\hat{\text{Cov}}(\tilde{b})]_i - [\text{Cov}(\tilde{b})]_i}{[\text{Cov}(\tilde{b})]_i} \right\}^{\frac{1}{2}} \quad (5.5.12)$$

where $[\text{Cov}(\tilde{b})]_i$ is the i^{th} element of the diagonal of $\text{Cov}(\tilde{b})$. This statistic is an average measure of percent deviations of estimated from actual standard errors. By comparing

diagonal elements of estimated and actual variance-covariance matrices, any such deviation in the covariance terms is being ignored.²

In Table 5.8 below, we report the average PRMSE of the estimators $\hat{\sigma}^2(X'X)^{-1}$, $\text{Cov}(b^*)$, and $\text{Cov}(\delta^*)$ for each value of R^2 .

It can be seen from Table 5.8 that the overall performance of the bootstrap in measuring the standard errors of least squares estimator is slightly worse (1.8%) than the usual estimator $\text{Cov}(b) = \hat{\sigma}^2(X'X)^{-1}$. The comparison of interest is that between columns 3 and 1.

Table 5.8
PRMSE of Estimated Covariance Matrices

R^2	$\hat{\sigma}^2(X'X)^{-1}$ PRMSE	$\text{Cov}(b^*)$ PRMSE	$\text{Cov}(\delta^*)$ PRMSE
0.00001	.2414	.2593	.4261
0.0100	.2414	.2593	.4182
0.0250	.2414	.2593	.4020
0.0500	.2414	.2593	.3749
0.0750	.2414	.2593	.3486
0.1000	.2414	.2593	.3255
0.2500	.2414	.2593	.2492
0.5000	.2414	.2593	.2203
0.7500	.2414	.2593	.2177
0.9000	.2414	.2593	.2168

Note that PRMSE $\text{Cov}(\delta^*)$ is considerably higher than that associated with $\hat{\sigma}^2(X'X)^{-1}$ for small values of R^2 . This is to be expected given that the bootstrap is overestimating the standard error of δ_i near the origin. As we move away from the origin, however, the degree of overestimation declines and the PRMSE associated with $\text{Cov}(\delta^*)$ declines as

well. For large values of R^2 , the bootstrap estimator of $\text{diag}[\text{Cov}(\delta)]$ is actually better than either m.v.u. or bootstrap estimator of $\text{diag}[\text{Cov}(b)]$.

5.5.3 Estimating Ellipsoids and Intervals

Having obtained estimates of $\text{Cov}(b)$ and $\text{Cov}(\delta)$, we can now derive estimators of confidence ellipsoids based on Q_1 , Q_2 , and Q_3 . In this section, estimators of Q_1 , Q_2 , and Q_3 are presented and their statistical properties, if known, are discussed.

Different estimators of A in (5.4.1) will quite naturally lead to different small sample distributions of Q_i ($i=1,2,3$) and therefore to different critical values for each ellipsoid. The usual estimator of $[\text{Cov}(b)]$ is $\hat{\sigma}^2(X'X)^{-1}$. Since $b \sim N(\beta, \sigma^2(X'X)^{-1})$ and $(T-K)\hat{\sigma}^2/\sigma^2 \sim \chi_K^2$ and is independent of b , we obtain the ellipsoid

$$Q_1(b) = (b-\hat{\beta})'X'X(b-\hat{\beta})/K\hat{\sigma}^2 \sim F_{K,T-K} \quad (5.5.13)$$

An alternative to $Q_1(b)$ which is centered at the least squares estimator is

$$Q_1(b^*) = (b-\hat{\beta})'[\text{Cov}(b^*)]^{-1}(b-\hat{\beta})/K \sim c F_{K,N-K} \quad (5.5.14)$$

where $c = \sigma^2(N-1)/(\hat{\sigma}^2(N-K))$. Unfortunately, the exact sampling distribution of this ellipsoid is conditional on $\hat{\sigma}^2$ and depends on the unknown parameter σ^2 . But,

$$T \rightarrow \infty \quad \text{plim}(c) = (N-1)/(N-K)$$

and consistent ellipsoids may be formed based on the $F_{K,N-K}$ distribution. Hence, $c^{-1}Q_1(b^*) \sim F_{K,N-K}$. This proposition

is tested using a chi-square goodness-of-fit test.³ The statistic is 41.41, indicating that the hypothesis cannot be rejected at any reasonable level of significance. If the proposition is true, then intervals based on $Q_1(b^*)$ may be formed using

$$b_i \pm t_{\alpha/2} b_i^{\sigma*} \quad (5.5.15)$$

where $t_{\alpha/2}$ is the critical value determined at the $1-\alpha/2$ level of significance from the t_{N-1} distribution,

$b_i^{\sigma*} = \{ \sum_{n=1}^N (b_{ni}^* - \bar{b}^*)^2 / (N-1) \}^{1/2}$ and $\bar{b}^* = \sum_{n=1}^N b_{ni}^* / N$. Thus, $b_i^{\sigma*}$ is the i^{th} diagonal element of $[\text{Cov}(b^*)]$.

Another ellipsoid similar to $Q_1(b)$ but which is centered at the James-Stein estimator has been suggested by Ullah, Carter and Srivastava [(UCS), 1984]. The UCS estimator is based on Q_2 and uses the m.v.u. estimator $\hat{\sigma}^2(X'X)^{-1}$ as an estimate of $\text{Cov}(b)$. Thus, define

$$Q_2(b) = (\delta - \beta)' X' X (\delta - \beta) / K \hat{\sigma}^2. \quad (5.5.16)$$

Since $\|\delta\| \leq \|b\|$, then $Q_2(b) \leq Q_1(b)$ and the ellipsoid centered at the James-Stein estimator will be smaller than the one centered at the MLE. Ullah, Carter, and Srivastava refer to this ratio as the "Improved-F".

Ullah, Carter, and Srivastava (1984) and Carter, Srivastava, Srivastava, and Ullah [CSSU, (1987)] justify ellipsoids of the form $Q_2(b)$ based on the fact that asymptotically, $\text{Cov}(\delta) = \text{Cov}(b)$; that is, $\lim \text{Cov}(\delta) = \text{Cov}(b)$. By the same reasoning it could be argued that as $T \rightarrow \infty$,

$$\lim (\delta - \beta)' [\text{Cov}(\delta)]^{-1} (\delta - \beta) / K = Q_1(b)$$

and the asymptotic test (as $T \rightarrow \infty$) based on use of the James-Stein rule is in fact the usual asymptotic test; such an extension leads to no apparent improvement and the ellipsoid $Q_2(b)$ is centered at the James-Stein estimator.

Ullah, Carter, and Srivastava (1984) and Carter, Srivastava, Srivastava, and Ullah [CSSU, (1987)] have studied the approximate sampling distribution of Q_2 where $[Cov(b)] = \hat{\sigma}^2(X'X)^{-1}$ using small- σ expansions and find it to be a weighted sum of central-F distributions. However, the use of their result to obtain critical points from the distribution of $Q_2(b)$ requires the researcher to speculate on the value of the quadratic form $\sigma^2/\beta'X'X\beta$. In other words, in order to use the UCS result, one must have some prior notion as to the value of the parameters in order to derive a confidence ellipsoid.

Although ellipsoids centered at δ but measured in $[Cov(\delta)]$ are smaller than those like $Q_2(b)$, estimating $Cov(\delta)$ by replacing β and σ^2 in (5.3.6) with statistics results in a covariance with an unknown and very complicated sampling distribution.

Ellipsoids of the form represented by Q_3 can be obtained by using $Cov(\delta) = [Cov(\delta^*)]$. Hence, the estimated ellipsoid becomes

$$Q_3(\delta^*) = (\delta - \beta)' [Cov(\delta^*)]^{-1} (\delta - \beta) / K \quad (5.5.17)$$

which has unknown small sample distribution.

Chi and Judge (1985) have constructed approximate 95% confidence intervals based on $Q_3(\delta^*)$ by using

$$\delta_i \pm z_{\alpha/2} \delta_i^* \quad (5.5.18)$$

where $z_{\alpha/2}$ is the critical value determined at the $\alpha/2$ level of significance for the $N(0,1)$ distribution. By using (5.5.18), Chi and Judge are implicitly arguing that

$Q_3(\delta^*) \rightarrow Q_1(b)$ asymptotically and $Q_3(\delta^*) \rightarrow \chi_K^2/K$. Although they find this approach performs reasonably well, it is likely that intervals centered at the James-Stein estimator can be improved since the justification for using (5.5.18) is derived solely from its asymptotic properties.

Although an exact confidence ellipsoid can be obtained using $Q_1(b)$, ones based on $Q_1(b^*)$, $Q_2(b)$, and $Q_3(\delta^*)$ cannot. This difficulty arises because the sampling distributions of these ellipsoids are unknown. The percentile bootstrap can once again be used to derive critical points for each of these quadratic forms. This procedure is described below.

5.5.4 Obtaining Critical Points

Although the exact sampling distribution of $Q_1(b)$ is known, those of the random ellipsoids $Q_1(b^*)$, $Q_2(b)$, and $Q_3(\delta^*)$ are uncertain. Given that the distribution of the quadratic form is unknown, one cannot form exact or approximate $100(1-\alpha)\%$ confidence intervals as in the case of $Q_1(b)$. The percentile bootstrap can be used to generate $100(1-\alpha)\%$ points from each of the empirical distributions of the quadratic forms considered above.

In the most general case, we can let $g(b):R^K \rightarrow R$ denote

a function of the random variable b whose distribution on the real line may or may not be known. The percentile bootstrap can be used to learn something about the distribution of the statistic $g(b)$. To do so, compute $g(b^*)$ for each of the N bootstrap estimates b^* forming the sequence $\{g(b^*)\}_1^N$. The N values of $\{g(b^*)\}$ are ranked in ascending order forming a bootstrap histogram. The $100(\alpha/2)\%$ and $100(1-\alpha)\%$ points from the histogram form the approximate $1-\alpha$ central confidence interval

$$g(b) \in [g(b^*(\alpha/2)), g(b^*(1-\alpha/2))] \quad (5.5.19)$$

for the function $g(b)$.

For ellipsoids centered at the least squares estimator there are two ways to generate critical points using the percentile bootstrap. First, since $b^* \sim N(b, \hat{\sigma}^2(X'X)^{-1})$ and $(T-K)\hat{\sigma}^{*2}/\hat{\sigma}^2 \sim \chi^2_{(T-K)}$, where $\hat{\sigma}^{*2} = (y^* - Xb^*)'(y^* - Xb^*)/(T-K)$, then

$$Q_1^*(b) = (b^* - b)'X'X(b^* - b)/\hat{\sigma}^{*2} \sim F_{K, T-K}. \quad (5.5.20)$$

Now, take a large number N of bootstrap samples $\{y^*\}_1^N$, calculate $\{Q_1^*(b)\}_1^N$ and form the empirical c.d.f. of $\hat{F}(n) = \#\{Q_1^* < n\}/N$. Inverting the empirical c.d.f. at the $(1-\alpha)\%$ point to get $Q_1^*(b, 1-\alpha)$ yields the approximate $100(1-\alpha)\%$ confidence ellipsoid

$$[0, Q_1^*(b, 1-\alpha)]. \quad (5.5.21)$$

Since the exact distribution of $Q_1(b)$ is known, we can compare the theoretical $100(1-\alpha)\%$ points from the $F_{K, T-K}$ distribution with those obtained using the percentile bootstrap. The average value of $Q_1^*(b^*)$ at $\alpha=.1$ was 1.92 and for $\alpha=.05$ it was 2.34. The theoretical critical values

from the $F_{K,T-K}$ distribution for these values of α are 1.97 and 2.40, respectively. The percentile bootstrap underestimates the exact critical values by a small amount.

A chi-square goodness-of-fit test was conducted to see whether the null hypothesis $H_0: Q_1^*(b) \sim F_{K,T-K}$ can be rejected.⁴ The test was performed using $\{Q_1^*(b)\}$ from 5 randomly selected draws of the Monte Carlo. The null hypothesis is rejected twice at the 5% level.⁵ In Figure 5.8 the histogram of a single sequence $\{Q_1^*(b)\}$ is superimposed on the histogram obtained from an $F_{K,T-K}$ random variable. Notice that the two are very similar.

The $100(1-\alpha)\%$ confidence points for $Q_1(b^*)$ can be obtained similarly to those for $Q_1(b)$, replacing $X'X/\hat{\sigma}^{*2}$ in (5.5.20) with $[\text{Cov}(b^*)]^{-1}$. That is,

$$Q_1^*(b^*) = (b^* - b)' [\text{Cov}(b^*)]^{-1} (b^* - b) / K \sim c_1 F_{K,N-K} \quad (5.5.22)$$

where $c_1 = (N-1)/(N-K)$.⁶ Thus, the exact sampling distribution of $Q_1^*(b^*)$ is known while that of $Q_1(b^*)$ is only approximately known. This yields the intriguing situation where the bootstrap critical point can be obtained from the usual set of tables; the use the percentile method is in principle unnecessary; However, the performance of percentile critical points for $Q_1(b^*)$ is of interest since the same principle is used below to derive approximate confidence ellipsoids centered at the James-Stein estimator. If the percentile bootstrap using $Q_1^*(b^*)$ fails to yield critical points which enable us to form ellipsoids of approximate coverage probability for

$Q_1(b^*)$, then we cannot expect $Q_3(\delta^*)$ to perform well either.

Bootstrap $100(1-\alpha)\%$ confidence ellipsoids can be obtained for $Q_2(b)$, and $Q_3(\delta^*)$ in an analogous fashion. The difference is that b^* is everywhere replaced by the bootstrap Stein estimates δ^* and the inverse covariance $[\text{Cov}(b)]^{-1}$ by the inverse of the appropriate estimated covariance matrix.

In the Monte Carlo experiment several ways of obtaining $100(1-\alpha)\%$ critical points for ellipsoids centered at the Stein estimates δ are explored. For the ellipsoid $Q_2(b)$, we generate the empirical c.d.f.'s in the way described above using the N order statistics from the bootstrap statistic

$$Q_2^*(b) = (\delta^* - b)' X' X (\delta^* - b) / K \hat{\sigma}^2. \quad (5.5.23)$$

It should be noted that $Q_2^*(b) \rightarrow \chi_K^2 / K$.

For the ellipsoid Q_3 , we can use the order statistics from the bootstrap statistic

$$Q_3^*(\delta^*) = (\delta^* - b)' [\text{Cov}(\delta^*)]^{-1} (\delta^* - b) / K \quad (5.5.24)$$

and the percentile method to obtain approximate $100(1-\alpha)\%$ critical points.⁷

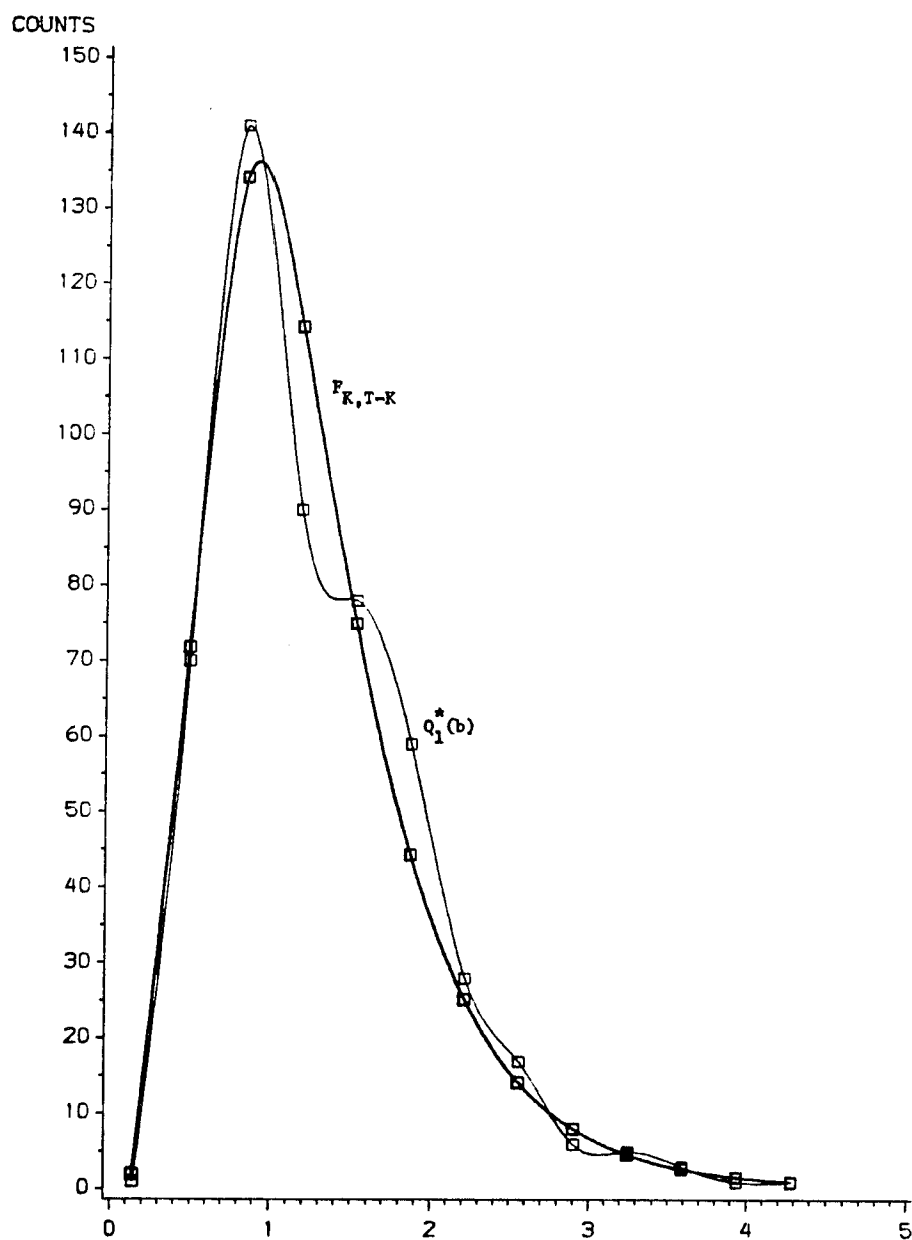
HISTOGRAMS OF $Q_1^*(b)$ AND $F(K,T-K)$ RANDOM VARIABLES

FIGURE 5.8

5.5.5 Summary

In the next section, the coverage probability and size of each of the following ellipsoids is considered:

$$Q_1(b) = (b - \hat{\beta})' X' X (b - \hat{\beta}) / K \hat{\sigma}^2$$

$$Q_1(b^*) = (b - \hat{\beta})' [\text{Cov}(b^*)]^{-1} (b - \hat{\beta}) / K$$

$$Q_2(b) = (\delta - \hat{\beta})' X' X (\delta - \hat{\beta}) / K \hat{\sigma}^2$$

$$Q_3(\delta^*) = (\delta - \hat{\beta})' [\text{Cov}(\delta^*)]^{-1} (\delta - \hat{\beta}) / K.$$

The coverage probability and size of the ellipsoids $Q_1(b)$ and $Q_1(b^*)$ are computed based on the 90% and 95% critical points generated from the percentile bootstrap using

$$Q_1^*(b) = (b^* - b)' X' X (b^* - b) / \hat{\sigma}^{*2} \quad (5.5.20)$$

and

$$Q_1^*(b^*) = (b^* - b)' [\text{Cov}(b^*)]^{-1} (b^* - b) / K \quad (5.5.19)$$

respectively.

The coverage probability and size of the ellipsoid $Q_2(b)$ is computed based on the 90% and 95% critical points generated from the percentile bootstrap using

$$Q_2^*(b) = (\delta^* - b)' X' X (\delta^* - b) / K \hat{\sigma}^{*2}. \quad (5.5.23)$$

Finally, the coverage probabilities and size of the ellipsoid $Q_3(\delta^*)$ is computed using 90% and 95% percentile bootstrap critical values obtained from

$$Q_3^*(\delta^*) = (\delta^* - b)' [\text{Cov}(\delta^*)]^{-1} (\delta^* - b) / K \quad (5.5.24)$$

In addition, bootstrap intervals based on $Q_1(b^*)$ and $Q_3(\delta^*)$ are compared to one another using:

$$b_i \pm t_{\alpha/2} b_i^{\sigma*} \quad (5.5.15)$$

$$\delta_i \pm z_{\alpha/2} \delta_i^{\sigma*}. \quad (5.5.18)$$

5.6 Results

In the following two subsection, the coverage probabilities and volumes of the various confidence procedures are compared.

5.6.1 Intervals

The results for 90% and 95% confidence intervals based on the use of (5.5.15) and (5.5.18) are reported in Tables 5.9 and 5.10, respectively. These intervals are similar in many respects to the percentile intervals described above in Tables 5.6 and 5.7. Since the number of bootstraps N is large, the critical points for (5.5.15) are taken from the $N(0,1)$ distribution [i.e., $N-1=499>120$] and these intervals are referred to as those obtained using the "normal approximation." As before, the LS bootstrap intervals are slightly too short to cover at the nominal 90% and 95% levels. The nominal 90% intervals (Table 5.9) centered at least squares have an average length of 3.2 and cover 88% of the time. The 95% intervals (Table 5.10) have an average length of 3.82 and cover slightly less than 94% of the time. Again, the underestimation of standard error by the bootstrap is likely to be responsible for this discrepancy.

From Table 5.9, it can be seen that at the origin the 90% intervals based on the normal approximation and centered at the James-Stein estimates (5.5.18) are 28% shorter than similar intervals centered at LS. In addition, these intervals cover 96% to 94% of the time for

$R^2 = .01$ to $.05$. Given the overestimation of the standard error of δ_1 by the bootstrap for small R^2 , it is not surprising that these intervals are too large. That is, their length could be further reduced without forcing the coverage probability below its nominal level.

At an R^2 of $.05$ the average length of these intervals has increased to 2.46 , which is about 77% the length of LS intervals. Coverage probability drops below the 90% level for $R^2 = .25$ and reaches a minimum of 84% at $R^2 = .9$. Once again, the biased intervals and ellipsoids are expected to perform poorly as bias increases. Hence, for the James-Stein estimator the test statistic used to control the degree of shrinkage is failing to distinguish true from false nonsample information a large proportion of the time; the intervals, which are on average centered in the wrong place, are too small to cover the true point with the expected frequency. It is expected that coverage will increase as $R^2 \rightarrow 1$ and JS estimates intervals converge to LS intervals. Recall, however, that the bootstrap is actually underestimating the standard error of δ_1 at high values of R^2 ($.5-.9$) and as a consequence it will yield confidence intervals which are too small in that region of the parameter space.

The 95% intervals for James-Stein estimates perform in similar fashion. Although the coverage probability drops as R^2 rises, it remains above the 95% level at $R^2 = .10$. Here, intervals are approximately 77% as long as those

centered at LS estimates. Coverage is minimized at an $R^2 = .5$ and begins to increase as the size and center of James-Stein intervals approaches that of least squares.

The percentile intervals centered at the Stein estimates are slightly smaller than those based on the normal approximation (5.5.18) and coverages for small R^2 (i.e., .00001, .01, .025) are close to the nominal level for 90% and 95% ellipsoids. However, we conclude that intervals centered at the JS estimates and based on the normal approximation (5.5.18) are preferred to percentile intervals (at least in the orthonormal model) since they perform better when R^2 is between 0.075 and .75. An interesting fact about the percentile intervals centered at LS and those like (5.5.15) is that the two are very similar in terms of size and coverage. The percentile method is known to work well for statistics with symmetric distributions and Efron (1986) has suggested improvements for percentile intervals when the distribution is skewed. Perhaps these methods could be used to improve the performance of bootstrap confidence intervals centered at the James-Stein estimates (5.5.18) for intermediate values of R^2 .

Table 5.9
Normal Approximation Intervals
90%

	R^2	.00001	.010	.025	.050	.075	.10	.25	.50	.75	.90
b_1 coverage		0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89
b_1 length		3.18	3.18	3.18	3.18	3.18	3.18	3.18	3.18	3.18	3.18
δ_1 coverage		0.97	0.96	0.96	0.95	0.94	0.94	0.87	0.85	0.85	0.86
δ_1 length		2.29	2.31	2.34	2.38	2.41	2.45	2.60	2.76	2.85	2.89
b_2 coverage		0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87
b_2 length		3.22	3.22	3.22	3.22	3.22	3.22	3.22	3.22	3.22	3.22
δ_2 coverage		0.95	0.95	0.94	0.93	0.92	0.91	0.86	0.83	0.83	0.84
δ_2 length		2.34	2.36	2.39	2.43	2.46	2.50	2.65	2.81	2.90	2.94
b_3 coverage		0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86
b_3 length		3.22	3.22	3.22	3.22	3.22	3.22	3.22	3.22	3.22	3.22
δ_3 coverage		0.95	0.95	0.94	0.93	0.91	0.90	0.85	0.84	0.85	0.85
δ_3 length		2.34	2.35	2.37	2.41	2.44	2.48	2.63	2.80	2.89	2.93
b_4 coverage		0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90
b_4 length		3.21	3.21	3.21	3.21	3.21	3.21	3.21	3.21	3.21	3.21
δ_4 coverage		0.96	0.95	0.94	0.94	0.94	0.92	0.87	0.87	0.88	0.88
δ_4 length		2.32	2.34	2.37	2.41	2.45	2.48	2.64	2.80	2.89	2.93
b_5 coverage		0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88
b_5 length		3.20	3.20	3.20	3.20	3.20	3.20	3.20	3.20	3.20	3.20
δ_5 coverage		0.95	0.95	0.95	0.95	0.94	0.94	0.88	0.87	0.87	0.87
δ_5 length		2.32	2.34	2.37	2.41	2.44	2.48	2.64	2.80	2.89	2.92
b_6 coverage		0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87
b_6 length		3.17	3.17	3.17	3.17	3.17	3.17	3.17	3.17	3.17	3.17
δ_6 coverage		0.95	0.96	0.95	0.94	0.94	0.93	0.87	0.85	0.84	0.85
δ_6 length		2.30	2.32	2.34	2.38	2.42	2.45	2.60	2.76	2.85	2.89
b_7 coverage		0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89
b_7 length		3.19	3.19	3.19	3.19	3.19	3.19	3.19	3.19	3.19	3.19
δ_7 coverage		0.95	0.96	0.96	0.96	0.95	0.94	0.88	0.87	0.87	0.88
δ_7 length		2.30	2.33	2.35	2.39	2.43	2.46	2.62	2.78	2.87	2.91
b_8 coverage		0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87
b_8 length		3.24	3.24	3.24	3.24	3.24	3.24	3.24	3.24	3.24	3.24
δ_8 coverage		0.94	0.95	0.95	0.94	0.92	0.92	0.86	0.84	0.84	0.84
δ_8 length		2.36	2.38	2.40	2.44	2.48	2.51	2.66	2.82	2.91	2.95

Table 5.10
Normal Approximation Intervals
95%

	R^2	.00001	.010	.025	.050	.075	.10	.25	.50	.75	.90
b_1 coverage		0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
b_1 length		3.79	3.79	3.79	3.79	3.79	3.79	3.79	3.79	3.79	3.79
δ_1 coverage		0.98	0.98	0.97	0.97	0.97	0.97	0.95	0.92	0.93	0.93
δ_1 length		2.73	2.76	2.79	2.83	2.88	2.92	3.10	3.29	3.40	3.45
b_2 coverage		0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92
b_2 length		3.84	3.84	3.84	3.84	3.84	3.84	3.84	3.84	3.84	3.84
δ_2 coverage		0.97	0.97	0.97	0.97	0.97	0.96	0.93	0.92	0.93	0.93
δ_2 length		2.79	2.81	2.85	2.89	2.94	2.97	3.16	3.35	3.45	3.50
b_3 coverage		0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
b_3 length		3.83	3.83	3.83	3.83	3.83	3.83	3.83	3.83	3.83	3.83
δ_3 coverage		0.98	0.98	0.98	0.97	0.97	0.95	0.93	0.90	0.91	0.91
δ_3 length		2.78	2.80	2.83	2.87	2.91	2.95	3.14	3.33	3.45	3.49
b_4 coverage		0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
b_4 length		3.83	3.83	3.83	3.83	3.83	3.83	3.83	3.83	3.83	3.83
δ_4 coverage		0.98	0.98	0.97	0.97	0.96	0.97	0.94	0.93	0.93	0.93
δ_4 length		2.77	2.79	2.82	2.87	2.92	2.96	3.15	3.34	3.45	3.49
b_5 coverage		0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
b_5 length		3.81	3.81	3.81	3.81	3.81	3.81	3.81	3.81	3.81	3.81
δ_5 coverage		0.97	0.98	0.97	0.97	0.97	0.96	0.94	0.93	0.93	0.93
δ_5 length		2.76	2.79	2.82	2.87	2.91	2.95	3.15	3.33	3.44	3.48
b_6 coverage		0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
b_6 length		3.78	3.78	3.78	3.78	3.78	3.78	3.78	3.78	3.78	3.78
δ_6 coverage		0.98	0.98	0.98	0.97	0.97	0.96	0.93	0.91	0.91	0.91
δ_6 length		2.74	2.77	2.79	2.84	2.88	2.92	3.10	3.28	3.39	3.44
b_7 coverage		0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
b_7 length		3.80	3.80	3.80	3.80	3.80	3.80	3.80	3.80	3.80	3.80
δ_7 coverage		0.97	0.97	0.97	0.97	0.97	0.96	0.94	0.92	0.93	0.93
δ_7 length		2.74	2.77	2.80	2.85	2.90	2.94	3.13	3.31	3.42	3.47
b_8 coverage		0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
b_8 length		3.86	3.86	3.86	3.86	3.86	3.86	3.86	3.86	3.86	3.86
δ_8 coverage		0.98	0.97	0.97	0.97	0.96	0.95	0.93	0.91	0.91	0.92
δ_8 length		2.81	2.84	2.86	2.91	2.95	2.99	3.17	3.36	3.47	3.52

5.6.2 Ellipsoids

From Table 5.11 it can be seen that the ellipsoid $Q_1(b)$ with critical point chosen using the percentile bootstrap with $Q_1^*(b)$ covers 87% of the time and has an average volume of 21.29. The ellipsoid based on $Q_1(b^*)$ with 90% critical point chosen using $Q_1^*(b^*)$ covers the true point only 80% of the time. This ellipsoid is considerably smaller than $Q_1(b)$ i.e., 10.9. The coverage probability and size of ellipsoids centered at least squares are independent of the true location parameters. Once again, underestimation of the standard error by the bootstrap will tend to cause these ellipsoids to be too small to cover at their nominal level. For $\text{Cov}(b^*)$ the situation is worse since low variability will not only cause the bootstrap value of c^* to be too small, but will also result in ellipsoidal semiaxes which are shorter than they otherwise would be.

The 95% ellipsoids (Table 5.12) centered at LS perform slightly better than the 90% ellipsoids. The ellipsoid $Q_1(b)$ with critical point chosen using the 95% point from the empirical distribution of the bootstrap statistic $Q_1^*(b)$ covers 94% of the time and has an average volume of 46.35. The ellipsoid based on $Q_1(b^*)$ with 95% critical point chosen using $Q_1^*(b^*)$ covers the true point just 88% of the time. As before, this ellipsoid is smaller than $Q_1(b)$.

The coverage probability and size of $100(1-\alpha)\%$ confidence ellipsoids centered at the James-Stein estimates

will vary as we move away from the origin. In Table 5.11, we see that the computed 90% ellipsoids based on the improved-F (i.e., Q_2) perform rather well. Using $Q_2^*(b)$ to obtain approximate 90% critical points resulted in ellipsoids which cover 96% of the time for $R^2 = .00001$ to 0.1. Coverage drops to 93% at $R^2 = .25$ and only falls below the nominal level of 90% at $R^2 = .9$. At the origin, the ellipsoid has an average volume of 4.12 which is 25% that of LS. As $R^2 \rightarrow .9$, volume approaches 13.02. In terms of volume and coverage probability, the use of $Q_2(b)$ with the percentile bootstrap is very encouraging.

For 95% ellipsoids (Table 5.12) of the same type the procedure again works quite well. At the origin, nominal 95% intervals cover 98% of the time. Coverage drops to 97% at $R^2 = .25$ and never falls below .95. Near the origin the average volume of these ellipsoids is $1/4^{\text{th}}$ that of similar ellipsoids centered at LS and are 61% the size of $Q_1(b)$ at $R^2 = .9$.

Ellipsoids centered at the James-Stein estimator and measured in $\text{Cov}(\delta^*)$ perform less spectacularly. This outcome might have been anticipated given the lackluster performance of the similarly derived ellipsoids $Q_1(b^*)$. However, these ellipsoids are considerably smaller than those based on Q_1 and Q_2 and cover above the nominal level for $R^2 \leq .10$. In Table 5.11 it can be seen that the coverage of $Q_3(\delta^*)$ is .99 at the origin and has only 8% of the volume of the usual ellipsoid $Q_1(b)$. Coverage of the 90%

ellipsoid falls to 96% at $R^2=.10$ and is 87%, 81%, and 80% as R^2 increases to .25, .50, and .75, respectively. Like the intervals centered at James-Stein estimates, the center and size of these ellipsoids depend on test statistic used to control the degree of shrinkage; the statistic fails to distinguish true from false nonsample information a large proportion of the time; hence, the ellipsoids are centered in the wrong place, on average, and are too small in certain directions of the parameter space to cover the true point with the expected frequency. At $R^2=.9$, the size of this ellipsoid has increased by a factor of 3.5 to 5.21 and covers only 80% of the time.

The performance of 95% ellipsoids (Table 5.12) is similar in that they cover 99% of the time from $R^2=.00001$ to .10 and then cover 98% and 92% for $R^2=.10$ and .25. The ellipsoid covers 88% of the time for $R^2=.5$, .75, and .90. The volume of this ellipsoid increases from 3.56 at the origin to 7.99 when $R^2=.9$.

The ellipsoids based on Q_3 represent significant improvements over those based on Q_1 and Q_2 near the origin, but their performance diminishes as bias increases. As mentioned before, this fact is not surprising given the relatively poor performance of $Q_1(b^*)$. In addition, we suspect that the performance of all the bootstrap ellipsoids considered will improve if the number of bootstrap replications is increased. In fact, there appears to be some evidence of this for $R^2=.00001$.⁸

Table 5.11
Percentile Confidence Ellipsoids
90%

	R^2	.00001	.010	.025	.050	.075	.10	.25	.50	.75	.90
$Q_1(b)$	coverage	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87
	volume	21.29	21.29	21.29	21.29	21.29	21.29	21.29	21.29	21.29	21.29
$Q_1(b^*)$	coverage	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86
	volume	10.90	10.90	10.90	10.90	10.90	10.90	10.90	10.90	10.90	10.90
$Q_2(b)$	coverage	0.96	0.96	0.96	0.96	0.96	0.96	0.93	0.91	0.90	0.89
	volume	4.12	4.27	4.51	4.88	5.24	5.59	7.52	10.24	12.13	13.02
$Q_3(\delta^*)$	coverage	0.99	0.99	0.99	0.98	0.97	0.96	0.87	0.81	0.80	0.80
	volume	1.45	1.51	1.58	1.69	1.81	1.92	2.62	3.73	4.70	5.21

Table 5.12
Percentile Confidence Ellipsoids
95%

	R^2	.00001	.010	.025	.050	.075	.10	.25	.50	.75	.90
$Q_1(b)$	coverage	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
	volume	46.35	46.35	46.35	46.35	46.35	46.35	46.35	46.35	46.35	46.35
$Q_1(b^*)$	coverage	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88
	volume	19.30	19.30	19.30	19.30	19.30	19.30	19.30	19.30	19.30	19.30
$Q_2(b)$	coverage	0.98	0.98	0.98	0.98	0.98	0.98	0.97	0.95	0.96	0.96
	volume	10.80	11.15	11.62	12.35	13.03	13.76	17.63	22.88	27.02	28.83
$Q_3(\delta^*)$	coverage	0.99	1.00	0.99	0.99	0.99	0.98	0.92	0.88	0.88	0.88
	volume	3.40	3.50	3.62	3.80	3.98	4.16	5.25	6.94	8.53	9.36

5.8 Conclusion

The bootstrap is able to provide approximate 90% and 95% confidence ellipsoids and intervals centered at the LS estimates for the orthonormal CNLRM. The percentile method and the method based on the normal approximation are remarkably similar to one another. The basic fault of the bootstrap in these instances is that it tends to underestimate standard error and consequently the ellipsoids and intervals tend to be too small to cover at their nominal levels. This feature has been noted by others and Swanepoel (1986) has suggested that this can be corrected by choosing bootstrap samples which contain fewer than T observations.

For bootstrap confidence intervals and ellipsoids centered at the James-Stein estimator, the results are not quite as certain. The main problem is that the bootstrap overestimates the standard error of the James-Stein estimator for small values of R^2 , causing intervals and ellipsoid to be too large. This difficulty is reversed for large values of R^2 . A satisfactory confidence procedure is one in which the true point is covered at or above the nominal level. Hence, the bootstrap provides satisfactory confidence intervals and ellipsoids over some regions of the parameter space, but not for others.

Presumably, one uses uncertain prior information when such information is believed to be reasonably good. That is, one uses the James-Stein estimator when one suspects

that the true value of β is zero. Because the risk improvements of the James-Stein estimator are greatest near the origin, this area of the parameter space is considered to be "more important" than the rest. All of the bootstrap procedures examined above work well in these regions of the parameter space. Not only do they cover the true point with greater probability, they are considerably smaller in size. In this sense, they represent significant improvements over classical methods.

¹Although the OLS residuals have been normalized, they remain linearly dependent. It may be appropriate to use residuals which are independent, like Theil's BLUS residuals. Like Freedman and Peters (1984) we recognize this point, but do not pursue it in this chapter.

²The following point is worth noting. As mentioned above, the RMSE criterion ignores covariance between elements \tilde{b}_i and \tilde{b}_j . Two other measures of covariance similarity have been computed:

$$M^{-1} \sum_1^M \det[\Sigma(\tilde{b})^{-1} \hat{\Sigma}(\tilde{b})]$$

$$M^{-1} \sum_1^M \text{trace}[\Sigma(\tilde{b})^{-1} \hat{\Sigma}(\tilde{b})]$$

where $\Sigma(\tilde{b})$ and $\hat{\Sigma}(\tilde{b})$ represent actual and estimated covariance matrices, respectively. In principle, if $\Sigma(\tilde{b}) = \hat{\Sigma}(\tilde{b})$, then these measures become $\det(I_K) = 1$ and $\text{trace}(I_K) = K$.

For the experiment above, $\Sigma(b) = I_K$ and we can concentrate on the expected value of $\det[\hat{\Sigma}(b)]$ versus that of $\det[\hat{\Sigma}(b^*)]$, which can be recognized as the sample generalized variance [Anderson (1984), pp. 259-265]. First, note that $E\{\det[\hat{\Sigma}(b)]\} = E[X_{T-K}^2 / T-K] = 8.23$. For the bootstrap estimator, $E\{\det[\hat{\Sigma}(b^*)]\} = E\{\det[W_K(N-1, \hat{\Sigma}(b))]\}$. Anderson [(1984), pp. 264-265] provides a simple formula for determining the value of this expectation. Using this result, we obtain

$$E\{\det[\hat{\Sigma}(b^*)]\} = (N-1)^{-K} \det[\hat{\Sigma}(b)] \prod_{i=1}^K (N-i) = 7.77$$

Given this outcome we report the average value of $\det[\hat{\Sigma}(b)]$ for least squares and the bootstrap:

$$M^{-1} \sum_{i=1}^M \det[\hat{\Sigma}(b)] = 5.30$$

$$M^{-1} \sum_{i=1}^M \det[\hat{\Sigma}(b^*)] = 5.16$$

and conclude that some discrepancy between actual and expected is occurring.

The other possibility is to use

$$M^{-1} \sum_{i=1}^M \text{trace}[\hat{\Sigma}(b)] = 7.84$$

$$M^{-1} \sum_{i=1}^M \text{trace}[\hat{\Sigma}(b^*)] = 7.80$$

where the expectation of each is equal to 8. By this measure we can see that the usual estimator of covariance performs slightly better than the bootstrap estimator. Given the lack of any distributional results for the covariance of the Stein-Rule, further discussion of the bootstrap estimate of its covariance matrix is omitted.

³Pearson's chi-square goodness-of-fit test [see DeGroot (1986), pp. 520-524] is used to determine whether the ellipsoids under consideration are distributed as approximate F or χ^2 random variables. To conduct these tests, the values of the bootstrap statistics $\{Q^*\}$ are ranked in ascending order $Q_1^* \leq \dots \leq Q_N^*$. The range $(Q_N^* - Q_1^*)$ is

divided into 40 intervals of equal length. Under the null hypothesis that the statistics from each sample represent drawings from a given distribution, the chi-square statistic is asymptotically χ^2_{39} . The 5% critical value for this test is 55.

⁴Ibid.

⁵A chi-square test of the hypothesis $H_0: Q_1^*(b) \sim F_{K,t-K}$ was conducted with the following results:

statistic	p-value
69.08	.002
28.24	.90
68.15	.002
37.77	.52
44.75	.24

⁶A chi-square test of the hypothesis $H_0: Q_1^*(b^*) \sim F_{K,N-K}$ was conducted with the following results:

statistic	p-value
60.7	.01
39.22	.46
66.12	.004
38.07	.51
40.07	.42

⁷In order to obtain critical points for the ellipsoid Q_3 , we could also use

$$Q_3^*(b) = (\delta^* - b) [\text{Cov}(\delta) |_{\mathbf{b}}]^{-1} (\delta^* - b) / K$$

where $\text{Cov}(\delta) |_{\mathbf{b}}$ is obtained from (5.3.6) by replacing the unknown parameters δ and σ^2 by b and $\hat{\sigma}^2$, respectively. The noncentrality parameter is obtained using the stein-rule estimates in place of δ . This last ellipsoid can be considered because the bootstrap allows us to treat b and $\hat{\sigma}^2$ as true parameter values when deriving the empirical

distribution of the statistic of interest. Unfortunately, for $R^2 \geq .5$ the estimator fails to yield positive definite covariance matrix estimates 100% of the time. We decided not to use it based on this troublesome fact. Below, we note the partial success of this estimate.

R^2	.00001	.01	.025	.05	.075	.10	.25
PRMSE	.1440	.1422	.1398	.1350	.1309	.1279	.1337
90% Ellipsoids							
coverage	.94	.95	.95	.94	.94	.94	.93
volume	2.10	2.17	2.28	2.42	2.56	2.68	3.37
95% Ellipsoids							
coverage	.96	.96	.96	.96	.96	.96	.95
volume	4.59	4.67	4.79	5.06	5.26	5.44	6.55

⁸Increasing the number of bootstrap samples to 1000 increases the accuracy of the approximate $100(1-\alpha)\%$ confidence ellipsoids. For the 90% ellipsoids the following changes occur:

$Q_1(b)$ from .87 to .89
 $Q_1(b^*)$ from .80 to .82
 $Q_2(b)$ from .96 to .995 ($R^2 = .00001$)
 $Q_3(\delta^*)$ no change ($R^2 = .00001$).

For the 95% ellipsoids the following changes occur:

$Q_1(b)$ from .945 to .945
 $Q_1(b^*)$ from .875 to .885
 $Q_2(b)$ from .985 to .987 ($R^2 = .00001$)
 $Q_3(\delta^*)$ from .997 to .995 ($R^2 = .00001$).

Chapter 6
Risk Characteristics of a Stein-Like Estimator for
the Probit Regression Model

- 6.1 Introduction
- 6.2 Classical Normal Linear Regression Model and Estimators
- 6.3 The Probit Regression Model
- 6.4 Shrinkage Estimator for the Probit Regression Model
- 6.5 Data Generation
- 6.6 Results
- 6.7 Conclusion

Chapter 6

Risk Characteristics of a Stein-Like Estimator for the Probit Regression Model

6.1 Introduction

For the classical normal linear regression model, the Stein-rule estimator dominates the maximum likelihood estimator if the number of hypothesis restrictions exceeds 2 and if certain other design related conditions hold [see Chapter 2]. To date, the theory of Stein-rule estimation has not been extended to include nonlinear statistical models. Despite the lack of an analytical result establishing a dominance property similar to the one for shrinking parameter estimates in linear models, the idea that risk improvements under squared error loss can be obtained in nonlinear models is certainly reasonable. In this chapter, a Stein-like shrinkage estimator of the probit regression model is proposed and its risk properties are studied in a Monte Carlo experiment.

There are many types of shrinkage estimators of the parameters of the classical normal linear regression model (CNLRM) linear statistical model which are known to dominate the MLE under quadratic loss [see Mittelhammer and Young (1981), Mittelhammer (1984); Stein (1961); Efron and Morris (1973); Judge and Bock (1978); and Vinod and Ullah (1983) to name only a few]. The basic principle in Stein estimation is to shrink maximum likelihood estimates of the model's parameters toward a set of restricted MLE's based on the value of the statistic used to test the hypothesis

restrictions. The proposed shrinkage estimator for probit models considered below works in exactly the same way and is based on a general family of minimax estimators proposed for use in linear statistical models by Mittelhammer and Young (1981) and extended by Mittelhammer (1984).

It is not known whether shrinkage can lead to risk improvements in nonlinear models like the probit. The difficulty arises from the fact that the exact sampling distribution of the MLE of the probit model, like that of nearly all nonlinear models, is unknown; inference in nonlinear models is usually based on the asymptotic normality of the MLE's [see Amemiya (1985), Ch. 4]. We will exploit the asymptotic normality of the MLE of the probit regression model to construct an analogue shrinkage estimator for its parameters.

The probit model is chosen for two reasons. First, the properties of the likelihood function are well-known and understood, i.e., its probability density is regular and concave, and the probit MLE's are distributed asymptotically normal. Second, the MLE's can be interpreted as the result of iterative generalized least squares [Amemiya (1985) and Finney (1952)]. These features of the probit model immediately suggest an algorithm for constructing a Stein-like estimator. First, obtain the maximum likelihood parameter estimates of the probit model. Then test the hypothesis restrictions using a likelihood ratio test. Finally, use this statistic to control the

shrinkage of the unconstrained MLE's toward the hypothesized values.

It is in this spirit that Dagenais (1985) has extended ridge regression to nonlinear models. As justification he notes that the nonlinear least squares estimator can be interpreted as weighted least squares and as such can be used iteratively in the usual ridge procedure. Schafer, Roi, and Wolfe (1984) use a similar method to explore the statistical properties of a ridge logistic estimator. Also, Copas (1983) has suggested a shrinkage estimator for binary regression models, but has not studied its risk properties.

In this chapter, the risk properties of a Stein-like estimator of the parameters of the probit model are studied in a Monte Carlo experiment. In our experiment, risk functions defined under squared error loss will be computed and compared using restricted, unrestricted, preliminary test, and shrinkage estimators for various degrees of hypothesis error.

The chapter is organized as follows. In section 6.2 we present the linear model and its estimators. In section 6.3 the probit model and its MLE are defined. In section 6.4 an analogue of the shrinkage estimator for linear models developed in 6.2 is proposed for the probit model. In section 6.5 the design of the experiment is discussed and in 6.6 the results are presented.

6.2 Classical Normal Linear Regression Model and Estimators

Before considering the probit model, we examine the classical normal linear regression model (CNLRM) and several of its estimators. In the following section we will develop analogues of these estimators for the probit model.

The CNLRM is represented by

$$y = X\beta + e \quad e \sim N(0, \sigma^2 I_T) \quad (6.2.1)$$

where y is a $T \times 1$ vector of observable random variables, X is a nonstochastic $T \times K$ matrix of rank K , β is a $K \times 1$ vector of unknown parameters, and e is a $T \times 1$ vector of unobservable normally and independently distributed random variables having zero mean and finite variance. The maximum likelihood estimator of β is

$$b = (X'X)^{-1}X'y = S^{-1}X'y. \quad (6.2.2)$$

The minimum variance unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = (y - Xb)'(y - Xb) / (T - K). \quad (6.2.3)$$

Now suppose that we have nonsample information in the form of exact linear hypothesis restrictions which we are willing to impose upon the model (6.2.1) in order to increase estimator precision. Let these restrictions be denoted $R\beta = r$, where the matrix R is known, $J \times K$, and of rank J and r is a $J \times 1$ vector of constants. The maximum likelihood estimator of β in (6.2.1) subject to $R\beta = r$ is

$$b_r = b - S^{-1}R'(RS^{-1}R')^{-1}(Rb - r). \quad (6.2.4)$$

Imposing restrictions of the form $R\beta = r$ on the estimation of the classical normal linear regression model

increases efficiency, but may induce bias if the hypotheses implied by the restrictions are not exactly true.

Another estimator which is commonly used is the preliminary test or pretest estimator [Bock, Yancy, and Judge (1973)]. With the pretest estimator, the hypothesis restrictions are tested using the sample. If they are rejected, then unrestricted estimates are used. If the restrictions cannot be rejected, then the nonsample information is imposed upon the model and the restricted MLE's are used. Formally, the pretest estimator is written

$$b_{PT} = I_{[0,c)}(u) b_r + I_{[c,\infty)}(u) b$$

where c is the $100(1-\alpha)\%$ critical value from the probability distribution associated with the test statistic u and I is the indicator function.

In deciding whether to use the restricted MLE, the unrestricted MLE, or the pretest estimator one would prefer to use a rule which ensures that on average the benefits obtained from greater efficiency outweigh the costs of using a biased estimator. This situation has given rise to important measures of an estimator's performance which explicitly take into account the costs of having to estimate parameters [see Chapter 2]. One such measure is risk under quadratic loss, which is a function of the distance of the estimator from the true parameter point. A loss function can be used to measure the cost of having to use an estimate of an unknown state of nature as the basis for making a decision. Thus, let \hat{b} be an arbitrary

estimator of an unknown vector β and let W be a positive definite and symmetric matrix. Weighted squared error loss is defined to be

$$L(\beta, \tilde{b}; W) = (\tilde{b} - \beta)' W (\tilde{b} - \beta) \quad (6.2.5)$$

and the associated risk of using \tilde{b} to estimate β is defined as

$$E[L(\beta, \tilde{b}; W)] = E[(\tilde{b} - \beta)' W (\tilde{b} - \beta)] = R(\beta, \tilde{b}; W). \quad (6.2.6)$$

Risk, then, is merely the average loss of having to use \tilde{b} as an estimator of β .

The risk of using the MLE in linear models is constant for all values of β . The risk of using the restricted MLE lies below that of the MLE for some value of β and above it for others. In fact, the risk of the restricted MLE is unbounded and increases as the degree of hypothesis error increases.

Although the risk of using the pretest estimator in the linear model is bounded, it is not minimax. Its risk under quadratic loss is below that of the MLE and above that of the restricted MLE for small degrees of hypothesis error. As hypothesis error increases, the risk of the pretest estimator rises above that of the MLE, reaches a maximum, and then falls, converging asymptotically from above to the risk associated with use of the MLE. And, as mentioned above in Chapter 2, the risk of using the pretest estimator is a function of the significance level of the hypothesis test. For most degrees of hypothesis error,

risk is inversely related to the size of the test.

Mittelhammer and Young (1981) and Mittelhammer (1984) have proposed a family of estimators which dominates the MLE of β in the CNLRM under weighted quadratic loss. Mittelhammer's estimator is Stein-like [James and Stein (1961)] and takes a convex combination of the unrestricted and restricted MLE's using

$$\delta = [1-c/u](b-b_r) + b_r \quad (6.2.8)$$

where

$u = (Rb-r)'(RS^{-1}R')^{-1}(Rb-r)/J\hat{\sigma}^2 \sim F_{J,T-K,\lambda}$, i.e., u is the conventional F-statistic used to test the hypothesis restrictions $H_0: R\beta=r$,

$$\lambda = (R\hat{\beta}-r)'(RS^{-1}R')^{-1}(R\hat{\beta}-r)/2\sigma^2,$$

$$c = a(T-K)/J,$$

$$0 < a < [2/(T-K+2)]\{\eta_L^{-1}\text{tr}[(RS^{-1}R')^{-1}RS^{-1}WS^{-1}R']-2\},$$

and η_L is the largest characteristic root of $[(RS^{-1}R')^{-1}RS^{-1}WS^{-1}R']$. The value of the constant 'a' which minimizes quadratic risk is found at the midpoint of the interval $[0, a_{\max}]$.

In general, (6.2.8) is dominated by a rather simple modification referred to as the positive-part rule. The positive-part rule prevents shrinkage of the OLS estimates beyond the restricted values, i.e., it prevents the Stein-rule from reversing the signs of the OLS estimates. Thus, if $u < c$, then the positive-part rule, denoted δ^+ , sets $c=u$. In this instance, equation (6.2.8) becomes $\delta^+ = b_r$. It must be noted, however, that no single value of 'a' has been

found which minimizes the risk of using the positive-part rule.

In the absence of any other nonsample information, we begin with the hypothesis restriction $H_0: \beta=0$ (which reduces the Mittelhammer estimator to the James-Stein (1961) estimator) and take $W=I_K$ (i.e., squared error loss).

In the sections below, we will consider the probit regression model and its maximum likelihood estimator. In section 6.4, a shrinkage estimator which is analagous to the Mittelhammer estimator is constructed and in section 6.6, the risk of this estimator is compared to that of the MLE, restricted MLE, and pretest estimators.

6.3 The Probit Regression Model

If the dependent variable y in the linear model above can take on a small number of discrete values, then the use of b as an estimator of β is often unsuitable. In these instances, the use of linear models like (6.2.1) above can lead to predictions which are inconsistent with the underlying data generation process [see Fomby, Hill, Johnson (1984) Ch. 16]. In this study, we consider the member of the class of qualitative response models (sometimes referred to as limited dependent variable models) in which the dependent variable is observed in one of two possible states.

Qualitative response models are defined as regression models in which the dependent variable takes discrete values. When a single dependent variable takes the value

of 1 or zero the model is called the binary choice model and is defined to be

$$\Pr(y_t=1) = F(x_t'\beta) \quad t=1,2,\dots,T$$

where $\{y_t\}$ is a sequence of independent binary random variables taking the value 1 or zero, x_t is a known $K \times 1$ vector of explanatory variables associated with the t^{th} observation, β is a $K \times 1$ vector of unknown parameters, and F is a certain cumulative distribution function. Choosing F to be a c.d.f. ensures that $x_t'\beta$ is mapped onto the interval $[0,1]$.

The most common choice of F is the standard normal c.d.f.

$$F(x_t'\beta) = \Phi(x_t'\beta) = \int_{-\infty}^{x_t'\beta} (2\pi)^{-\frac{1}{2}} \exp\{-\frac{1}{2}r^2\} dr. \quad (6.3.1)$$

The probit function is the inverse of the $N(0,1)$ c.d.f. Although the logistic function is similar in many respects to the probit function, the choice between the two should not be made arbitrarily; the model selected should be that which is most consistent with the underlying data generation process (DGP). (For a summary of the DGP which gives rise to the probit regression model, refer to section 2.3 in Chapter 3 above.)

Given that the probit model is believed to be consistent with the underlying data generation process, maximum likelihood estimation of the model's parameters is a simple application of numerical optimization [see section 3.1 of Chapter 3].

To define the likelihood function associated with the

probit regression model, let $\{y_1, y_2, \dots, y_T\}$ be a random sample of T Bernoulli trials with parameter π . Then, the probability density function is denoted

$$f(y_t | \pi) = \pi^{y_t} (1-\pi)^{1-y_t} \quad y=1, 2, \dots, T.$$

If

$$\pi_t = \Pr(y_t=1) = F(x_t' \beta) = \Phi(x_t' \beta) \quad t=1, 2, \dots, T,$$

then the likelihood function is denoted

$$L = \prod_{t=1}^T \Phi(x_t' \beta)^{y_t} [1-\Phi(x_t' \beta)]^{1-y_t} \quad (6.3.2)$$

and the log-likelihood is

$$\ell = \ln L = \sum_{t=1}^T y_t \ln(\Phi_t) + (1-y_t) \ln(1-\Phi_t) \quad (6.3.3)$$

where $\Phi_t = \Phi(x_t' \beta)$. Differentiation of ℓ with respect to β yields the gradient vector

$$g(\beta) = \partial \ell / \partial \beta = \sum_{t=1}^T \{(y_t - \Phi_t) / [\Phi_t(1-\Phi_t)]\} \phi_t x_t \quad (6.3.4)$$

where ϕ_t is the standard normal p.d.f. evaluated at the argument $x_t' \beta$. Given $g(\tilde{\beta})=0$, a sufficient condition for a local maximum of ℓ is for the Hessian

$$H(\beta) = - \sum_{t=1}^T \phi_t \{ y_t [(x_t' \beta) \Phi_t + \phi_t] / (\Phi_t)^2 + (1-y_t) [\phi_t - (x_t' \beta)(1-\Phi_t)] / (1-\Phi_t)^2 \} x_t x_t' \quad (6.3.5)$$

to be negative definite when evaluated at the estimate $\tilde{\beta}$.

The global concavity of the likelihood function [see

Amemiya (1985), p. 273] permits a liberal choice of

starting values; those from the OLS estimator $b_0 = (X'X)^{-1}X'y$

are convenient and are often used as the first approximation. Evaluating (6.3.4) and (6.3.5) at the starting value b_0 using the Newton-Raphson nonlinear maximization algorithm [see Chapter 3] with step length 1 yields the first round estimate

$$b_1 = b_0 - H^{-1}(b_0)g(b_0). \quad (6.3.6)$$

The first round estimate is then used in (6.3.6) again to get a second round estimate and so on until convergence [see Chapter 2].

The method of scoring technique can be used in a similar way; the only difference is that the negative of the Hessian is replaced with the Information matrix. For regular densities,

$$I(\beta) = - E[\partial^2 \ell / \partial \beta \partial \beta'] = E[(\partial \ell / \partial \beta)(\partial \ell / \partial \beta)']$$

and for the probit model

$$I(\beta) = E \sum_{t=1}^T \{ (y_t - \hat{F}_t)^2 / [\hat{F}_t(1 - \hat{F}_t)]^2 \} \phi_t^2 x_t x_t'$$

$$I(\beta) = \sum_{t=1}^T \{ [\phi_t^2 / \hat{F}_t(1 - \hat{F}_t)] \} x_t x_t'.$$

Although the Newton-Raphson and the method of scoring lead to the same point estimates in the probit regression model (due to the global concavity of ℓ), each yields a distinct estimate of the asymptotic covariance matrix. Griffiths, Hill, and Pope (1987) investigate the small sample properties of the two methods for the probit model and conclude that they differ negligibly.

Using Nelder and Wedderburn's (1972) generalized linear model (GLIM), an identical set of parameter estimates for (6.2.1) can be obtained as a set iterated feasible generalized least squares estimates. The GLIM estimates are solutions to the maximum likelihood equations (6.3.6). In addition, GLIM estimates are obtained using the fact that for regular densities

$$I(\beta) = -E[\partial^2 \ell / \partial \beta \partial \beta'] = E[(\partial \ell / \partial \beta)(\partial \ell / \partial \beta)'].$$

Hence, they are equivalent to the MLE's obtained through the use of the method of scoring.

One of the reported advantages of GLIM estimation is that it requires the first derivatives only. However, caution is advised in interpreting the resulting estimate of the asymptotic covariance matrix. Although the Gauss-Newton, Newton-Raphson and the method of scoring algorithms yield identical parameter estimates, they will generally not yield identical estimates of the asymptotic covariance matrix. Thus, the estimated covariance of the GLIM estimator will differ from that obtained from the MLE using the Newton-Raphson, which is the algorithm used below.

The interpretation of the MLE's as a set of iterated feasible generalized least squares (IFGLS) estimates suggests an algorithm for constructing a Stein-like shrinkage estimator for the probit model; we simply substitute the IFGLS probit estimates in place of the OLS estimates in the usual Stein-rule estimator. Although Copas (1983) appeals to the generalized linear model

interpretation of the binary regression model to derive a similar shrinkage estimator of its parameters, the statistical properties of the resulting estimator are unknown. Difficulties arise from (1) having to use a statistic which is asymptotically distributed as an F random variable to control the degree of shrinkage, (2) the fact that MLE's are not normally distributed in small samples, and (3) the values of the shrinkage constant which insures minimaxity--if they exist--are unknown.

6.4 Shrinkage Estimator for the Probit Regression Model

The proposed shrinkage estimator for the probit regression model is similar in form to that used for the CNLRM and is similar to the one used in Copas (1983). In principle, each of the components of (6.2.8) must be replaced by similar statistics from the probit model. That is, we need the sets of unrestricted and restricted estimates, a test statistic, and a shrinkage constant (α).

The restricted and unrestricted MLE's of β in the probit model are easily obtained and are used in the proposed Stein-like estimator. We use the MLE (6.3.6) of β as our analogue of b in (6.2.8) and, under the null hypothesis $H_0: \beta=0$, the restricted estimate of β , denoted b_r , will be a $K \times 1$ vector of zeros.

Several alternatives are available for use as a test statistic to control the degree of shrinkage. The usual statistics are derived based on Lagrange multiplier (LM), likelihood ratio (LR), and Wald principles. Davidson and

MacKinnon have compared the performance of LM and LR in small samples using a Monte Carlo experiment and conclude that LM performs better under the null hypothesis. However, they also find that the LR test performs slightly better under the alternative hypothesis. Griffiths, Hill and Pope (1987) use the Wald statistic as the basis for pretesting and find that its small sample performance leaves much to be desired. This results because the power function associated with this particular statistic is nonmonotonic [see Nelson and Savin (1988) who show that on the other hand the power functions of the LR and LM statistics used in this and other nonlinear models are monotonic].

Since the unrestricted likelihood function must be obtained and because it is important that the asymptotic statistic used have small sample distribution close to the noncentral F, we use the likelihood ratio test of the null hypothesis $H_0: R\beta=r$, which has the form

$$LR = 2[\ell(b_n) - \ell(b_r)] \sim \chi^2_J \quad (\text{if } H_0 \text{ true})$$

where $\ell(b_r)$ and $\ell(b_n)$ represent the values of the log likelihood function evaluated at the restricted and unrestricted estimates, respectively. This statistic is further modified by dividing by the number of restrictions imposed, J , yielding

$$u^* = 2[\ell(b_n) - \ell(b_r)] \sim \chi^2_J/J \quad (6.4.1)$$

This statistic coincides asymptotically with u from (6.2.8) since $\lim u \rightarrow \chi^2_J/J$.

The shrinkage constant is obtained by replacing $(x'x)^{-1}=S^{-1}$ with the estimated covariance matrix from the probit model. Thus we choose $a(b_n)$ such that

$$0 \leq a(b_n) \leq [2/(T-K+2)] \{n_L^{-1} \text{tr}[H(b_n)^{-1}W] - 2\} \quad (6.4.2)$$

where n_L^{-1} is the largest characteristic root of $H(b_n)W$ and $H(b_n)$ is the Hessian matrix evaluated at the n^{th} round estimate.

In the absence of any other nonsample information, and without loss of generality, we have chosen to shrink the unrestricted maximum likelihood estimates toward the hypothesis restriction that $\beta=0$; thus, the proposed Stein-like estimator of the probit model becomes

$$\delta(b_n) = (1-c^*/u^*)b_n \quad (6.4.3)$$

where

$$u^* = 2[\ell(b_n) - \ell(b_r)] \sim \chi^2_J/J$$

b_n is the n^{th} round estimate from Newton-Raphson,

$$b_r = 0,$$

$$c^* = a(b_n) [T-K]/K$$

$$0 \leq a(b_n) \leq [2/(T-K+2)] \{n_L^{-1} \text{tr}[H(b_n)^{-1}W] - 2\}$$

$$K \geq 3, \text{ and}$$

$$W = I_K.$$

The empirical risk of the estimator (6.4.8) under squared error loss (i.e., $W=I_K$) is explored below in a Monte Carlo experiment where it is compared to that of the MLE, restricted MLE, and pretest estimators.

6.5 Data Generation

Initially, we choose two sample sizes ($T=50, 100$) and 8 orthonormal regressors (i.e., $K=8$ and $X'X=I_8$). The regressors must be chosen with some care since an experiment with too much variability in $x_t'\beta$ will result in values of π that accumulate in the tails of the c.d.f. This generally results in MLE's with large standard errors and the Newton-Raphson often fails to converge. The X matrix consists of positive values and β were chosen such that $|F(x_t'\beta)| < 2$. Thus, for this experiment we let $\beta = kL$, where

$$L = [1, 1, 1, 1, -1, -1, -1, -1]$$

and k is a constant chosen from the vector

$$k = [.00001, 0.1, 0.5, 1.0, 1.5, 2.0].$$

Recall from section 6.3 that given x_t and β , $\pi_t = F(x_t'\beta)$ is the probability that $y_t=1$. Thus, single value of the dependent variable y_t may be obtained by drawing a uniform random number $u_t \in [0,1]$ and putting

$$y_t = \begin{cases} 1 & \text{if } u_t \in [0, \pi_t] \\ 0 & \text{if } u_t \in (\pi_t, 1]. \end{cases}$$

We draw 1000 random samples of size $T=50$ and 100 for each of the 6 design points β using this method.

6.6 Results

The risk performance of Stein-like shrinkage estimators for the probit model is very encouraging. Below, we report the results of the positive-part rule and for nonpositive part rules where $a(b_n)$ takes on quartile values on the interval $[0, a_{\max}]$ with

$$a_{\max} = [2/(T-K+2)] \{n_L^{-1} \text{tr}[H(b_n)^{-1}] - 2\}.$$

For samples of size 50 note from Table 6.1 that the risk of using the MLE increases uniformly as $k \uparrow 2$; and, it is highly variable for all values of k (i.e., standard error of the empirical risk ranges from 13.09 at $k=.0001$ to 26.21 at $k=2.0$). In fact, the risk of the RMLE (b_r) crosses that of the MLE somewhere between $k=1.5$ and $k=2.0$. A sample of size 50 is apparently too small for asymptotic results to apply, a fact noted by Griffiths, Hill, and Pope (1987).

In Figure 6.1 we see that the risk of the pretest estimator is considerably lower than that of the MLE, and rises above that of the MLE at about $k=1.0$. The positive-part Stein-rule with $a(b_n)=a_{\max} (\delta^+)$ dominates the pretest estimator and the MLE. The risk of using the positive-part rule is greater than that of the RMLE for $k < 1.5$. The Stein-rule which takes $a(b_n)$ at the midpoint of the given interval is clearly the best of the nonpositive-part rules. Note, however, that its risk is greater than that of the positive-part rule for $k < 2.0$.

For reasonable degrees of hypothesis error, the

positive-part rule performs much as we would expect and certainly offers an improvement over the MLE for samples of this size. The risk gain of the positive-part rule at $k=.0001$ and $.10$ is over 800%. At $k=1.5$ and 2.0 , the relative risk improvement is 36% and 7%, respectively.

Finally, note that the Stein-rule rule which uses $a(b_n)=a_{\max}$ has risk of 27.34 for $k=2.0$, indicating that the MLE is not dominated. The minimax condition may have to be modified somewhat to ensure dominance of the Stein-rule over the MLE for samples of this size.

Table 6.1
Risk Characteristics
 $T=50$

k	.0001	.10	.50	1.0	1.5	2.0
MLE	18.72 (13.09)	18.77 (13.28)	19.36 (14.68)	20.77 (14.87)	22.73 (18.74)	26.17 (26.21)
b_r	0.00 (0.00)	0.08 (0.00)	2.00 (0.00)	8.00 (0.00)	18.00 (0.00)	32.00 (0.00)
b_{PT}	4.42 (14.79)	4.74 (15.12)	7.09 (16.21)	14.30 (15.43)	23.51 (17.06)	32.75 (24.36)
δ^+ $a=a_{\max}$	2.16 (6.44)	2.24 (6.65)	4.21 (7.99)	9.34 (7.51)	16.67 (12.09)	24.28 (20.20)
Stein, (a_{\max})	11.12 (19.70)	11.19 (19.52)	12.85 (19.24)	16.47 (17.89)	21.66 (17.80)	27.34 (23.79)
Stein, $3a_{\max}/4$	7.01 (11.69)	7.10 (11.69)	8.89 (12.43)	12.98 (12.05)	18.39 (14.97)	24.43 (22.49)
Stein, $a_{\max}/2$	6.91 (10.28)	7.00 (10.42)	8.66 (11.58)	12.54 (11.30)	17.47 (15.24)	23.26 (23.01)
Stein, $a_{\max}/4$	10.82 (12.12)	10.89 (12.28)	12.15 (13.44)	15.13 (13.15)	18.92 (16.94)	23.84 (24.48)
% Reject $\alpha=.05$.089	.096	.10	.19	.30	.48

(standard error of the risk in parentheses)

The performance of the MLE and other probit estimators for samples of size 100 is given in Table 6.2. Notice that the risk of using the MLE drops by 25% at $k=.0001$ and by 35% at $k=2.0$. Also, the risk of the MLE is considerably less variable than it was for $T=50$. The standard errors, which range between 13.09 and 26.21, still appear to be quite large.

In Figure 6.2 we see that for $T=100$, the risk of the RMLE crosses that of the MLE between $k=1.0$ and $k=1.5$, while that of the pretest ($\alpha=.05$) crosses that of the MLE between $k=0.5$ and $k=1.0$. Once again, the positive-part rule δ^+ performs extremely well, with risk of only .59 for $k=.0001$ and .68 for $k=.10$. These values represent extremely large percentage risk improvements over the MLE (i.e., 3000% and 2100%, respectively). For $k=1.5$ and 2.0 , the positive part rule offers 38% and 26% risk improvement over the MLE, respectively. The standard error of measurement is also fairly low, especially for $k=.0001-1.0$ where it is below 3.

The midpoint of the shrinkage interval is again the best of the remaining Stein-rule estimators. The risk of the rule using the midpoint is 4.48 for $k=.0001$ and only 13.77 for $k=2$. Thus, the MLE is over 3 times as risky at $k=.0001$ and 1.3 times as risky at $k=2.0$. In addition, note that the "minimaxity" condition appears to be sufficient since risk of each of the Stein-rules falls below that of the MLE for all values of k .

Table 6.2
Risk Characteristics
T=100

k	.0001	.10	.50	1.0	1.5	2.0
MLE	14.73 (8.77)	14.83 (8.81)	14.92 (8.60)	15.56 (9.03)	16.16 (10.06)	17.86 (11.65)
b_r	0.00 (0.00)	0.08 (0.00)	2.00 (0.00)	8.00 (0.00)	18.00 (0.00)	32.00 (0.00)
b_{PT}	2.72 (9.82)	2.72 (9.73)	5.26 (9.55)	13.02 (9.09)	19.05 (8.35)	20.42 (12.08)
δ^+ $a=a_{\max}$	0.59 (2.64)	0.68 (2.62)	2.45 (2.26)	7.12 (2.90)	11.63 (5.84)	14.16 (8.48)
Stein, a_{\max}	14.90 (27.72)	14.57 (25.54)	13.51 (18.84)	13.36 (14.20)	13.58 (9.95)	14.47 (9.45)
Stein, $3a_{\max}/4$	7.11 (14.28)	7.01 (13.00)	7.42 (9.46)	9.50 (7.95)	11.44 (7.27)	13.53 (8.56)
Stein, $a_{\max}/2$	4.48 (7.06)	4.53 (6.64)	5.62 (5.54)	8.60 (5.90)	11.15 (7.18)	13.77 (9.07)
Stein, $a_{\max}/4$	7.02 (7.57)	7.13 (7.59)	8.12 (7.16)	10.63 (7.38)	12.73 (8.55)	15.22 (10.29)
% Reject $\alpha=.05$.075	.0	.11	.30	.60	.87

(standard error of the risk in parentheses)

6.7 Conclusion

The risk performance of the shrinkage estimator for the probit model is similar in most respects to that of its linear model analogue. For T=50, the Stein-like estimator performs at least as well as the MLE. Unfortunately, this sample size appears to be too small to yield very precise estimates. It is doubtful whether the asymptotics hold and maximum likelihood estimation cannot be recommended.

For larger samples, the performance of each of the

estimators examined improves. The MLE is dominated by the Stein-like estimator; however, the positive-part rule does not dominate the nonpositive-part rule for large degrees of hypothesis error. Nevertheless, the positive part rule is clearly superior to the MLE and other Stein-rule alternatives for small to moderate degrees of hypothesis error.

Finally, it should not be surprising that the Stein-like estimator examined here appears to have properties similar to those of its linear analogue. If $b_n \sim N(\beta, I(\beta))$ and $u^* \sim F_{J, T-K}$, then it seems reasonable that the theory which is used to generate the minimaxity result in the linear model should carry through to the nonlinear model. Unfortunately, this remains to be shown with any kind of rigor. The first step would be a clear demonstration of the finite sample properties of the MLE's of the probit's parameters. Before this is done it is unlikely that analytical progress will be made in the more complicated case examined above.

RISK OF MLE, RMLE, PRETEST, STEIN, AND
POSITIVE-PART STEIN RULE ESTIMATORS

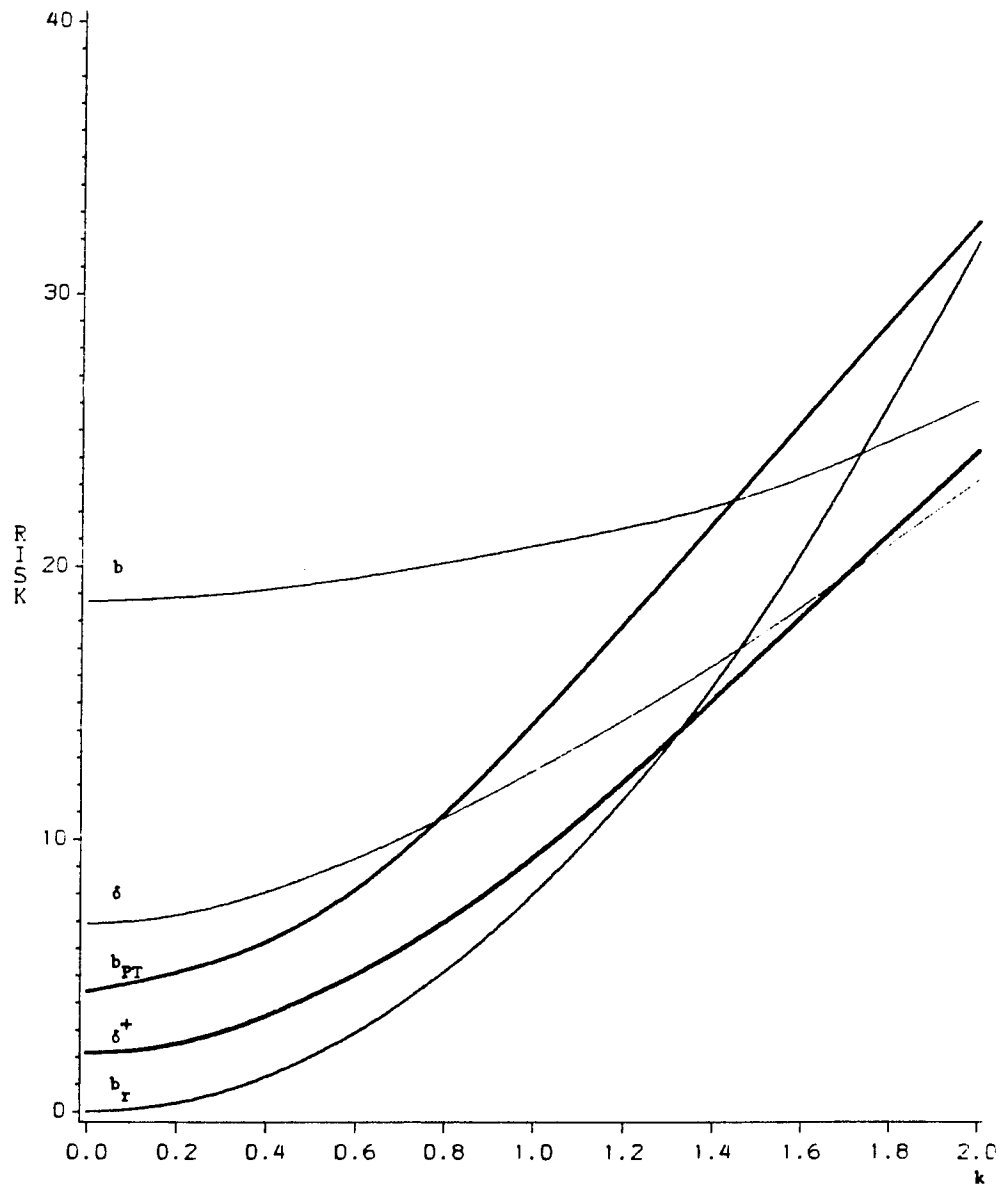


FIGURE 6.1

T=50

RISK OF MLE, RMLE, PRETEST, STEIN, AND
POSITIVE-PART STEIN RULE ESTIMATORS

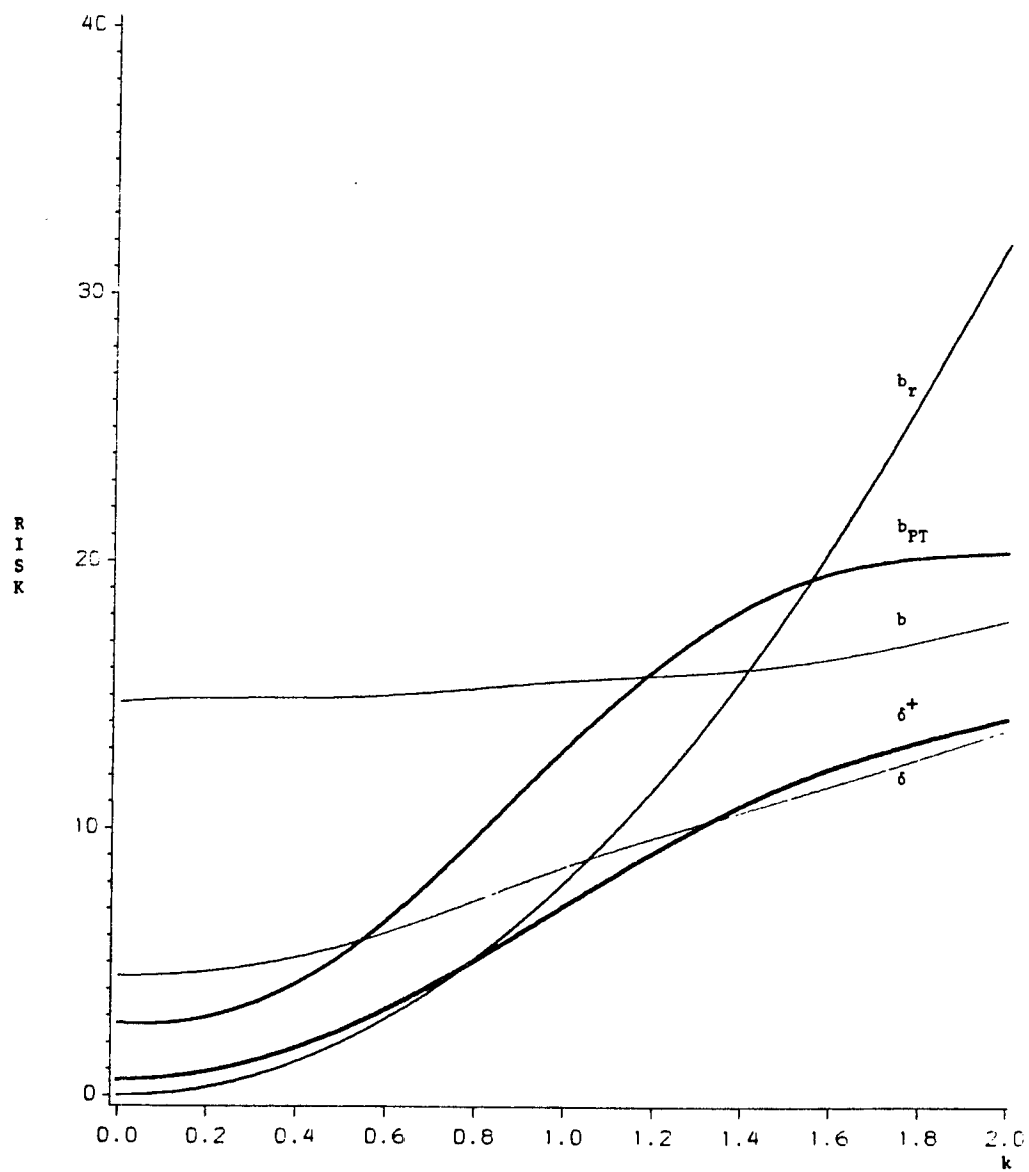


FIGURE 6.2

T=100

Chapter 7 Concluding Remarks

Three issues in Stein estimation have been explored in this dissertation: forecasting, confidence sets, and use in nonlinear models. In the first 3 chapters the important literature is surveyed, the basic results of Stein estimation are outlined, and related estimation procedures which are used in the later chapters are summarized. Chapters 4, 5 and 6 represent my contribution to the existing body of literature in Stein-rule estimation.

In Chapter 4, the Stein-rule is used in conjunction with a well-known macroeconomic model to generate out-of-sample forecasts of nominal GNP growth. These forecasts were compared to traditional ones generated by ordinary least squares, restricted least squares, pretest, and ARIMA estimators using a root-mean-square error criterion.

One of the shrinkage prediction rules examined in this chapter allows one to combine the mean forecast (i.e., the average value of the variable over the sample) with those of an explanatory model. The prior information employed by this rule is that future values of the time series will tend to be close to what they have been in the past, i.e., the future values will be equal to the past average. This is certainly reasonable if the time series is stationary. The resulting shrinkage forecasts are linear combinations of those from the explanatory model and the in-sample average of the time series. If the explanatory model is

not supported by the data, then the forecasted values of the time series will be close to the in-sample mean. If the model is well supported by the sample, then the forecasts are approximately equal to those generated by least squares.

In the example used in Chapter 4, the mean shrinkage forecast rule performed quite well, having lower RMSE than OLS, RLS, and pretest forecast rules over most forecast horizons. Although the mean shrinkage forecasting rule did not outperform the ARIMA model in terms of RMSE, it was argued that the shrinkage forecasts are better in another sense. The ARIMA will often fail to predict turning points in the time series which the explanatory model may pick up.

Other shrinkage rules were examined. Nearly all of these outperformed least squares and many outperformed their RLS counterparts. There is reason to believe that shrinkage predictors are indeed robust to differences between in- and out-of-sample data differences and additional effort should be made to extend the small body of existing results in this literature.

Several research topics can be pursued. First, it would be useful to know how Stein-rule forecasts compare to those generated from other kinds of forecasting models like vector autoregressions [Sims (1980)], Bayesian vector autoregressions [Litterman (1986)], multivariate ARIMA's, ridge regression, and even other Stein-like procedures like New-Stein [Stein (1981)], truncated empirical Bayes [Efron

and Morris (1972)], and generalized robust Bayes [Berger (1980)]. Another possibility is to examine the forecasting performance of rules which use the sample in an efficient way to determine what kind of restrictions to impose. The poor performance of the model selection/pretest shrinkage estimators (as well as the wide array of models and RMSE's they produce) suggests that improvements could be made. Judge, Hill and Bock (1988) are working at establishing rules which use the sample to select restrictions based on minimization of expected loss. These models could certainly be used to forecast in situations like those explored above.

There is a great deal of foundational work which needs to be done on the forecasting problem. Thanks to work by Copas (1983), Jones and Copas (1986) and Fomby and Hill (1988), we are just now beginning to understand how deviations between X and X_0 affect least squares and shrinkage prediction. In particular, the work of Fomby and Hill enable us to examine the effects of marginal deviation in mean, rotation, and variation of X and X_0 . This advance will enable useful simulations to be performed demonstrating the robustness of various types of estimators describe above.

Confidence intervals are also needed for biased forecast rules. Although many analytical results are within reach, it will be some time before much of the current work in advanced analysis can be readily used. The

bootstrap may provide an adequate alternative until these other approaches can be fully developed.

One of the serious drawbacks of estimators from the Stein family is that an acceptable means of assessing their precision has not yet been obtained. In Chapter 5 the bootstrap was used in various ways to obtain approximate $100(1-\alpha)\%$ confidence intervals and ellipsoids centered at the least squares and James-Stein estimators of the orthonormal CNLRM. It was shown that the bootstrap is able to provide approximate 90% and 95% confidence procedures centered at least squares. The basic shortcoming of the procedure is that it tends to underestimate standard error of LS and consequently leads to intervals and ellipsoids which are too small to cover at their nominal levels.

For bootstrap confidence procedures centered at the James-Stein estimator, the results were mixed. The main problem is that the bootstrap overestimates the standard error of the James-Stein estimator for small signal-to-noise ratios, causing intervals and ellipsoids to be too large. This difficulty was reversed for large signal-to-noise ratios. However, it was concluded that despite these difficulties, bootstrap confidence sets centered at the James-Stein estimator are considerably smaller and cover with higher probability than least squares confidence sets in important regions of the parameter space. Hence, they represent local improvements over existing procedures.

Many questions remain to be answered. For instance

- (1) Why does the bootstrap fail to get the correct standard error for the James-Stein estimator? What can be done to correct this and what are the consequences for confidence sets?
- (2) In general, how well does the bootstrap approximate an unknown covariance matrix? Under what circumstances can it be made better?
- (3) What effect would the use of BLUS or other independent residuals have on our results?
- (4) How do confidence intervals and ellipsoids perform when the data are collinear? What effect would different choices of β have on the results (i.e., all values of β_i not identical)?
- (5) How do empirical Bayes confidence sets compare to the bootstrap confidence sets when σ^2 is assumed unknown?
- (6) Can the bootstrap be used in conjunction with empirical Bayes approach to derive better confidence intervals and ellipsoids?
- (7) How well does the bootstrap work for more complicated Stein-like estimators like the componentwise Stein-rule [Stein 1981])?
- (8) Given that the confidence ellipsoid is sufficiently understood, how does one go about performing hypothesis tests of a given size and what are the power characteristics of the resulting tests?

The precision issue is important and difficult. The approach taken here and the suggestions for future research

listed above must be viewed as preliminary. Analytical advances will almost certainly be made based on the work of Phillips (1984) and other, but these seem far away. In the interim, empirically based procedures are competitors of the various approximation methods which have been suggested. In this respect, it would be useful to extend the scope of the small-sigma asymptotic expansions for the F-statistic proposed by Ullah, Carter and Srivastava (1984) and compare these results to those obtained using the bootstrap.

Finally, in Chapter 6 a shrinkage estimator of the parameters of the probit regression model was proposed and its properties studied in a Monte Carlo. The risk performance of the estimator (for samples of size 100) is similar in most respects to that of its linear model analogue. For samples of size 50, the MLE cannot be recommended since risk is above that of RMLE, pretest and Stein-like estimator for a very large portion of the feasible parameter space. For larger samples, the performance of the MLE improves and is dominated by the Stein-like estimator.

The results are preliminary, but promising. Many issues remain which need to be addressed. Some attempt must be made to understand the nature and control of collinearity in the probit model when the number of regressors is large. Until this is done, it will be difficult to design simulations which explore in a

systematic way the consequences of collinearity in probit models in general and for the shrinkage estimator in particular. Having solved this problem, a study should be conducted exploring the performance of this estimator using different parameter points and experimental design.

Other nonlinear estimators may also give way to shrinkage counterparts which have lower quadratic risk. These include nonlinear least squares estimators, Box-Cox estimators, maximum likelihood logit estimators, and many others.

References

- Abrahamse, A. P. J. and Koerts, J. (1969). "A Comparison Between the Power of the Durbin Watson Test and the Power of the BLUS Test." Journal of the American Statistical Association, 64, 938-948.
- Ahmed, E. and Johannes, J. M. (1984). "St. Louis Equation Restrictions and Criticisms Revisited." Journal of Money, Credit and Banking, 16, 514-520.
- Akaike, H. (1970). "Statistical Predictor Identification." The Annals of the Institute of Statistical Mathematics, 2, 203-217.
- Akaike, H. (1973). "Information Theory and an Extension of the Maximum Likelihood Principle." In B. N. Petrov and F. Csaki, eds. 2nd International Symposium on Information Theory. Budapest: Akademiai Kiado.
- Akaike, H. (1974). "A New Look at the Statistical Model Identification." IEEE Transactions on Automatic Control, AC-19, 716-723.
- Almon, S. (1965). "The Distributed Lag Between Capital Appropriations and Expenditures." Econometrica, 33, 178-196.
- Amemiya, T. (1985). Advanced Econometrics. Cambridge: Harvard University Press.
- Amemiya, T. (1980). "Selection of Regressors." International Economic Review, 21, 331-354.
- Andersen, L. C. and Jordan, J. L. (1968). "Monetary and Fiscal Actions: A Test of their Relative Importance in Economic Stabilization." Federal Reserve Bank of St. Louis Review, 50, 11-24.
- Anderson, R. L. (1942). "Distribution of the Serial Correlation Coefficient." Annals of Mathematical Statistics, 13, 1-33.
- Anderson, T. W. (1984). An Introduction to Multivariate Statistical Analysis, second edition. New York: John Wiley and Sons.
- Anderson, T. W. (1977). "Estimation for Autoregressive Moving Average Models in the Time and Frequency Domain." Annals of Statistics, 5, 842-865.
- Ansley, C. F. (1979). "An Algorithm for the Exact Likelihood of a Mixed Autoregressive-Moving Average

- Process." Biometrika, 66, 59-65.
- Baranchik, A. (1964). "Multiple Regression Estimation of the Mean of a Multivariate Normal Distribution." Technical Report No. 51. Department of Statistics, Stanford University.
- Bard, Y. (1974). Nonlinear Parameter Estimation. New York: Academic Press.
- Batten, D. S. and Hafer, R. W. (1983). "The relative Impact of Monetary and Fiscal Actions on Economic Activity: A Cross-Country Comparison." Federal Reserve Bank of St. Louis Review, 65, 5-12.
- Batten, D. S. and Thornton, D. C. (1984). "How Robust are the Policy Conclusions of the St. Louis Equation." Federal Reserve Bank of St. Louis Review, 66, 26-32.
- Batten, D. S. and Thornton, D. C. (1983). "Polynomial Distributed Lags and the Estimation of the St. Louis Equation." Federal Reserve Bank of St. Louis Review, 65, 13-25.
- Belsley, D. A. (1984). "Demeaning Conditioning Diagnostics Through Centering." American Statistician, 38, 73-77.
- Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). Regression Diagnostics: Identifying Influential Observations and Sources of Collinearity. New York: John Wiley and Sons, Inc.
- Berndt, E., Hall, B., Hall, R., and Hausman, J. (1974). "Estimation and Inference in Nonlinear Structural Models." Annals of Economic and Social Measurement, 3, 653-665.
- Beran, R. (1982). "Estimated Sampling Distributions: The Bootstrap and Its Competitors." Annals of Statistics, 10, 212-225.
- Berger, J. (1980a). "A Robust Generalized Bayes Estimator and Confidence Region for a Multivariate Normal Mean." Annals of Statistics, 8, 716-761.
- Berger, J. (1980b). Statistical Decision Theory: Foundations, Concepts, and Method. New York: Springer-Verlag.
- Berger, J. and Berliner, L. M. (1984). "Bayesian Input In Stein Estimation and a New Minimax Empirical Bayes Estimator." Journal of Econometrics, 25, 87-108.

- Berger, J. and Bock, M. E. (1975). "Minimax Estimation of a Multivariate Normal Mean With Arbitrary Quadratic Loss." Technical Report 143. Purdue University.
- Bickel, P. J. and Doksum, K. A. (1977). Mathematical Statistics: Basic Ideas and Selected Topics. Oakland, California: Holden-Day, Inc.
- Bickel, P. J. and Freedman, D. A. (1981). "Some Asymptotic Theory for the Bootstrap." Annals of Statistics, 9, 1196-1217.
- Bock, M. E., Yancy, T. A. and Judge, G. (1973). "The Statistical Consequences of Preliminary Test Estimators in Regression." Journal of the American Statistical Association, 68, 109-116.
- Box, G. E. P. and Cox, D. R. (1964). "An Analysis of Transformations." Journal of the Royal Statistical Society, series B., 26, 211-234.
- Box, G. E. P. and Jenkins, G. M. (1976). Time Series Analysis Forecasting and Control. Revised edition. Oakland, California: Holden-Day.
- Box, G. E. P. and Pierce, D. A. (1970). "Distribution of Residual Autocorrelations in Autoregressive Integrated Moving Average Time Series Models." Journal of the American Statistical Association, 65, 1509-1526.
- Bradley, E. (1973). "The Equivalence of Maximum Likelihood and Weighted Least Squares Estimates in the Exponential Family." Journal of the American Statistical Association, 68, 199-200.
- Breusch, T. S. and Pagan, A. R. (1979). "A Simple Test for Heteroscedasticity and Random Coefficient Models." Econometrica, 47, 1287-1294.
- Brook, R. J. (1976). "On the Use of a Regret Function to Set Significance Points in Prior Tests of Estimation." Journal of the American Statistical Association, 71, 126-131.
- Brown, L. D. (1966). "On the Admissibility of Invariant Estimators of One of More Location Parameters." Annals of Mathematical Statistics, 37, 1087-1136.
- Carlson, K. M. (1986). "Recent Revisions in GNP." Federal Reserve Bank of St. Louis Review, 68, 17-24.
- Carter, R. A. L., Srivastava, M. S., Srivastava, V. K., and Ullah, A. (1987). "Unbiased Estimation of the MSE Matrix of Stein-Rule Estimators, Confidence Ellipsoids

and Hypothesis Testing." Mimeo.

- Casella, G. (1980). "Minimax Ridge Regression Estimation." Annals of Statistics, 8, 1036-1056.
- Chatfield, C. (1984). The Analysis of Time Series: An Introduction. 3rd ed. New York: Chapman and Hall.
- Chi, X. W. and Judge G. C. (1985). "On Assessing the Precision of Stein's Estimator." Economic Letters, 18, 143-148.
- Cochrane, D. and Orcutt, G. H. (1949). "Application of Least Squares Regression to Relationships Containing Autocorrelated Error Terms." Journal of the American Statistical Association, 44, 32-61.
- Copas, J. B. (1983). "Regression, Prediction and Shrinkage." Journal of the Royal Statistical Society, series B, 45, 311-354.
- Dagenais, M. G. (1983). "Extension of the Ridge Regression Technique to Non-Linear Models with Additive Errors." Economic Letters, 12, 169-174.
- Davidson, R. and MacKinnon, J. G. (1984). "Convenient Specification Tests for Logit and Probit Models." Journal of Econometrics, 25, 241-262.
- DeGroot, M. H. (1970). Optimal Statistical Decisions. New York: McGraw-Hill.
- DeGroot, M. H. (1986). Probability and Statistics, Second Edition. Reading, Massachusetts: Addison-Wesley Publishing Company.
- DeLury, D. (1950). Values and Integrals of Orthogonal Polynomials Up to $n = 26$. Toronto: University of Toronto Press.
- Dempster, A. P., Schatzoff, M., and Wermuth, M. (1977). "A Simulation Study of Alternatives to Ordinary Least Squares." Journal of the American Statistical Association, 72, 77-104.
- Dey, D. K. and Berger, J. O. (1983). "On Trauncation of Shrinkage Estimators in Simultaneous Estimation of Normal Means." Journal of the American Statistical Association, 78, 865-869.
- Dey, D. K. and Srinivasan, C. (1985). "Estimation of a Covariance Matrix Under Stein's Loss." Annals of Statistics, 13, 1581-1591.

- Dhrymes, P. J. (1974). Econometrics: Statistical Foundations and Applications. New York: Springer-Verlag, Inc.
- Efron, B. (1987). "Better Bootstrap Confidence Intervals." Journal of the American Statistical Association, 82, 171-185.
- Efron, B. (1979). "Computers and the Theory of Statistics: Thinking the Unthinkable." SIAM Review, 21, 460-480.
- Efron, B. (1982). The Jackknife, the Bootstrap, and Other Resampling Plans. Philadelphia: Society for Industrial and Applied Mathematics.
- Efron, B. (1981). "Nonparametric Estimates of Standard Errors: The Jackknife, the Bootstrap, and Other Methods." Biometrika, 68, 589-599.
- Efron, B. and Morris, C. (1972). "Limiting the Risk of Bayes and Empirical Bayes Estimators--Part II: The Empirical Bayes Case." Journal of the American Statistical Association, 67, 130-139.
- Efron, B. and Morris, C. (1977). "Stein's Paradox in Statistics." Scientific American, 236, 119-127.
- Efron, B. and Morris, C. (1973). "Stein's Rule and Its Competitors--An Empirical Bayes Approach." Journal of the American Statistical Association, 68, 117-130.
- Engle, R. F. (1984). "Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics." In Handbook of Econometrics, Volume II. Edited by Z. Griliches and M. D. Intriligator. Amsterdam: North-Holland.
- Fair, R. C. (1986). "Evaluating the Predictive Accuracy of Models." In Handbook of Econometrics, Volume III. Edited by Z. Griliches and M. D. Intriligator. New York: Elsevier Science Publishers, B.V.
- Finney, D. J. (1952). Probit Analysis. Second edition. Cambridge: University Press.
- Fisher, F. (1970). "Tests on Equality Between Sets of Coefficients in Two Linear Regressions: An Expository Note." Econometrica, 28, 361-366.
- Fletcher, R. and Powell, M. J. D. (1963). "A Rapidly Convergent Descent Method for Minimization." The Computer Journal, 6, 163-168.

- Fomby, T. B. and Hill, R. C. (1988). "The Geometry of Least Squares Prediction." Mimeo.
- Fomby, T. B. and Hill, R. C. (1985). "The Relative Efficiency of a Robust Generalized Bayes Estimator in a Linear Regression Model with Multicollinearity." Mimeo.
- Fomby, T. B., Hill, R. C. and Johnson, S. R. (1985). Advanced Econometric Methods. New York: Springer-Verlag.
- Freedman, D. A. (1981). "Bootstrapping Regression Models." The Annals of Statistics, 9, 1218-1228.
- Freedman, D. A. and Peters, S. C. (1984). "Bootstrapping a Regression Equation: Some Empirical Results." Journal of the American Statistical Association, 79, 97-106.
- Friedman, L. W. and Friedman, H. H. (1984). "Statistical Considerations in Computer Simulation: The State of The Art." Journal of Statistical Computer Simulation, 19, 237-263.
- Gallant, R. A. (1975). "The Power of the Likelihood Ratio Test of Location in Nonlinear Regression Models." Journal of the American Statistical Association, 70, 198-203.
- Geweke, J. and Meese, R. (1981). "Estimating Regression Models of Finite But Unknown Order." International Economic Review, 22, 55-70.
- Gilbert, A. R. (1985). "Operating Procedures for Conducting Monetary Policy." Federal Reserve Bank of St. Louis Review, 67, 13-21.
- Goldfeld, S. M. and Quandt, R. E. (1972). Nonlinear Methods in Econometrics. Amsterdam: North-Holland Publishing Company.
- Goldfield, S. M., Quandt, R. E. and Trotter, H. F. (1966). "Maximization by Quadratic Hill-Climbing." Econometrica, 34, 541-551.
- Granger, C. W. J. and Newbold, P. (1977). Forecasting Economic Time Series. New York: Academic Press.
- Griffiths, W. E., Hill, R. C., and Pope, P. J. (1987). "An Investigation into the Small Sample Properties of Covariance Matrix and Pre-Test Estimators for the Probit Model." Journal of the American Statistical Association, 82, 929-937.

- Hafer S. and Hein S. (1980). "The Dynamics and Estimation of Short-Run Money Demand." Federal Reserve Board of St. Louis Review, 62, 26-35.
- Harvey, A. C. (1983a). The Econometric Analysis of Time Series. New York: John Wiley and Sons.
- Harvey, A. C. (1983b). Time Series Models. New York: John Wiley and Sons.
- Hammersley, J. M. and Handscomb, D. C. (1964). Monte Carlo Methods. New York: John Wiley and Sons, Inc.
- Hemmerle, W. J. and Brantle, T. F. (1978). "Explicit and Constrained Generalized Ridge Regression." Technometrics, 20, 109-120.
- Hendry, D. F. (1984). "Monte Carlo Experimentation in Econometrics." In Handbook of Econometrics Volume II. Edited by Z. Griliches and M. D. Intriligator. New York: Elsevier Science Publishers B.V.
- Hey, J. D. (1981). Disequilibrium in Economics. New York: New York University Press.
- Hill, R. C. (1982). "Restrictions Imposed by the Almon Polynomial Distributed Lag." Mimeo.
- Hill, R. C. and Fomby, T. B. (1986). "Improved Confidence Sets in a Non-Utopian Setting." In Advances in Econometrics Volume 5. Guest Edited by Daniel J. Slottje and edited by George F. Rhodes. Greenwich, Connecticut: JAI Press, Inc.
- Hill, R. C. and Judge, G. C. (1987). "Improved Prediction Using a Principal Components Estimator." Mimeo.
- Hoerl, A. E. and Kennard, R. W. (1970a). "Ridge Regression: Applications to Non-Orthogonal Problems." Technometrics, 12, 55-67.
- Hoerl, A. E. and Kennard, R. W. (1970b). "Ridge Regression: Biased Estimation for Non-Orthogonal Problems." Technometrics, 12, 69-82.
- Hoerl, A. E. and Kennard, R. W., and Baldwin, K. F. (1975). "Ridge Regression: Some Simulations." Communications in Statistics, A, 4, 105-123.
- Hogg, R. V. and Craig, A. T. (1978). Introduction to Mathematical Statistics. Fourth Edition. New York: MacMillan Publishing Company, Inc.

- Hopmans, A. C. M. and Kleijnen, J. P. C. (1980). "Regression Estimation in Simulation." Journal of the Operations Research Society, 31, 1033-1038.
- Hwang, J. T. and Casella, G. (1982). "Minimax Confidence Sets for the Mean of a Multivariate Normal Distribution." Annals of Statistics, 10, 868-881.
- James, W. and Stein, C. (1961). "Estimation With Quadratic Loss." Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 1. Berkeley: University of California Press, 361-379.
- Jenkins, G. M. and Watts, D. G. (1968). Spectral Analysis and Its Applications. San Francisco: Holden-Day.
- Jennrich, R. I. and Oman, S. D. (1986). "How Much Does Stein Estimation Help in Multiple Linear Regression." Technometrics, 28, 113-122.
- Jeyaratnam, S. (1985). "Minimax Volume Confidence Regions." Statistics and Probability Letters, 3, 307-308.
- Jones, M. C. and Copas, J. B. (1986). "On the Robustness of Shrinkage Predictors in Regression to Differences Between Past and Future Data." Journal of the Royal Statistical Society, B, 48, 223-237.
- Joshi, V. M. (1967). "Inadmissibility of the Usual Confidence Sets for the Mean of a Multivariate Normal Population." Annals of Mathematical Statistics, 38, 1868-1875.
- Judge, George G. and Bock, M. E. (1978). The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics. Studies in Mathematical and Managerial Economics. Volume 25. Edited by Henri Theil. Amsterdam: North-Holland Publishing Company.
- Judge, G. G., Griffiths, W. E., Hill, R. C., Lutkepohl, H. and Tsoung-Chao Lee. (1985). The Theory and Practice of Econometrics. Second edition. New York: John Wiley and Sons.
- Judge, G. C. Miyazaki, S., and Yancy, T. A. (1985). "Some Information on the Sampling Performance of Stein's New Estimator." Economic Letters, 19, 253-256.
- Kadane, J. B. (1971). "Comparison of k-Class Estimators When the Disturbances are Small." Econometrica, 39, 723-737.

- Kendall, M. G. (1957). A Course in Multivariate Analysis. New York: Hafner.
- Kendall, M. G. (1945). "On the Analysis of Oscillating Time Series." Journal of the Royal Statistical Society, 108, 93-129.
- Kleijnen, P. C. (1976). "Discrete Simulation: Types, Applications and Problems." Proceedings of 8th AICA Congress (Delft, Aug. 23-28). In Simulation of Systems. Edited by L. Dekker. New York: North-Holland, 31-38.
- Koenker, R. (1981). "A Note on Studentizing a Test for Heteroscedasticity." Journal of Econometrics, 17, 107-112.
- Koerts, J. and Abrahamse, A. P. J. (1971). "On the Power of the BLUS Procedure." Journal of the American Statistical Association, 63, 1227-1236.
- Lawless, J. F. and Wang, P. (1976). "A Simulation Study of Ridge and Other Regression Estimators." Communications in Statistics, A, 5, 307-323.
- Lindley, D. V. (1961). "Discussion of Professor Stein's Paper." Journal of the Royal Statistical Society, B, 31, 285-287.
- Lindley, D. V. (1977). "The Distinction Between Inference and Decision." Synthese, 36, 51-58.
- Litterman, R. (1986). "Forecasting with Bayesian Vector Autoregressions--Five Years of Experience." Journal of Business and Economic Statistics, 4, 25-38.
- McCullough, P. (1983). "Quasi-Likelihood Functions." Annals of Statistics, 11, 59-67.
- McCullough, P. and Nelder, J. A. (1983). Generalized Linear Models. New York: Chapman and Hall.
- McDonnald, G. C. and Galarneau, D. I. (1975). "A Monte Carlo Evaluation of Some Ridge-Type Estimators." Journal of the American Statistical Association, 70, 407-416.
- Mallows, C. L. (1973). "Some Comments on Cp." Technometrics, 15, 661-676.
- Marquardt, D. W. (1963). "An Algorithm for Least Squares Estimation of Nonlinear Parameters." Journal of the Society of Industrial Applied Mathematics, 11, 431-441.

- Marquardt, D. W. (1970). "Generalized Inverse, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation." Technometrics, 12, 591-612.
- Mihram, G. A. (1973). "On Antithetic Variates." Proceedings of the Summer Computer Simulation Conference, 91-95.
- Mittelhammer, R. C. (1984). "Quadratic Risk Domination of Restricted Least Squares Estimators via Stein-Ruled Auxiliary Constraints." Journal of Econometrics, 25, 289-303.
- Mittelhammer, R. C. and Young, D. L. (1981). "Mitigating the Effects of Multicollinearity Using Exact and Stochastic Restrictions: Reply." American Journal of Agricultural Economics, 63, 301-304.
- Modigliani, F. and Ando, A. (1976). "The Impact of Fiscal Action on Aggregate Income and the Monetarist Controversy: Theory and Evidence." In Monetarism, edited by Jerome W. Stein, pp. 17-42. Amsterdam: North-Holland.
- Morris, C. (1977). "Interval Estimation for Empirical Bayes Generalizations for Stein's Estimator." Rand Corporation. Rand Paper Series.
- Morris, C. (1983). "Parametric Empirical Bayes Inference: Theory and Applications." Journal of the American Statistical Association, 78, 47-55.
- Muirhead, R. (1982). Aspects of Multivariate Statistical Analysis. New York: John Wiley and Sons.
- Myazaki, S., Judge, G. and Yancy, T. A. (1986). "Estimation of Location Parameters Under Nonnormal Errors and Quadratic Loss." Journal of Business and Economic Statistics, 4, 263-268.
- Nelder, J. A. and Wedderburn, R. W. M. (1972). "Generalized Linear Models." Journal of the Royal Statistical Society, A, 135, 370-384.
- Nelson, F. D. and Savin, N. E. (1988). "The Nonmonotonicity of the Power Function of the Wald Test in Nonlinear Models." (University of Iowa, Department of Economics, Working Paper No. 88-7).
- Nelson, H. L. and Granger, C. W. J. (1979). "Experience With Using the Box-Cox Transformation When Forecasting Economic Time Series." Journal of Econometrics, 10, 57-69.

- Newbold, P. (1974). "The Exact Likelihood Function for a Mixed Autoregressive-Moving Average Process." Biometrika, 61, 423-426.
- Ohtani, K. (1986). "A Distribution Function of the F-Ratio When the Stein-Rule Estimator is Used in Place of the OLS Estimator." Economic Letters, 21, 257-260.
- Ohtani, K. and Toyoda, T. (1979). "Estimation of Regression Coefficients After a Preliminary Test for Heteroskedasticity." Journal of Econometrics, 12, 151-159.
- Olshen, R. A. (1977). Comment on "A Note on a Reformulation of the S-Method of Multiple-Comparison." by Scheffe, H. Journal of the American Statistical Association, 72, 144-146.
- Oman, S. D. (1982). "Contracting Towards Subspaces When Estimating the Mean of a Multivariate Normal Distribution." Journal of Multivariate Analysis, 12, 270-290.
- Oulton, N. (1981). "Aggregate Investment and Tobin's Q: The Evidence for Britain." Oxford Economic Papers, 33, 177-202.
- Oxley, L. T. and Roberts, C. J. (1982). "Pitfalls in the Application of the Cochrane-Orcutt Technique." Oxford Bulletin of Economics and Statistics, 44, 227-240.
- Pagano, M. and Hartley, M. J. (1981). "On Fitting Distributed Lag Models Subject to Polynomial Restrictions." Journal of Econometrics, 16, 171-198.
- Parzen, E. (1974). "Some Recent Advances in Time Series Modeling." IEEE Transactions on Automatic Control. AC-19, 723-730.
- Pearson, E. S. and Hartley, H. O. editors. (1972). Biometrika Tables for Statisticians, Volume 2, Second Edition. New York: Cambridge University Press.
- Phillips, P. C. B. (1984). "The Exact Distribution of the Stein-Rule Estimator." Journal of Econometrics, 25, 123-131.
- Prais, S. J. and Winsten, C. B. (1954). "Trend Estimators and Serial Correlation." Cowles Commission Discussion Paper no. 383. Chicago.
- Quenouille, M. H. (1949). "Approximate Tests of Correlation in Time Series." Journal of the Royal

Statistical Society, B, 68.

- Ramsey, J. B. (1969). "Tests for Specification Errors in Classical Linear Least Squares Regression Analysis." Journal of the Royal Statistical Society, B, 31, 350-371.
- Rao, C. R. (1973). Linear Statistical Inference and Its Applications. Second Edition. New York: John Wiley and Sons.
- Sargan, J. D. (1976). "Econometric Estimators and the Edgeworth Approximation." Econometrica, 44, 421-448.
- Sawa, T. and Hiromatsu, T. (1973). "Minimax Regret Significance Points for a Preliminary Test in Regression Analysis." Econometrica, 41, 1093-1101.
- Schaefer, R. L., Roi, L. D., and Wolfe, R. A. (1984). "A Ridge Logistic Estimator." Communications in Statistics, Theory and Methods, 13, 99-113.
- Scheffe, H. (1977). "A Note on a Reformulation of the S-Method of Multiple-Comparisons." Journal of the American Statistical Association, 72, 143-144.
- Scheffe, H. (1959). The Analysis of Variance. New York: John Wiley and Sons, Inc.
- Schenker, N. (1985). "Qualms About Bootstrap Confidence Intervals." Journal of the American Statistical Association, 80, 360-361.
- Schmidt, P. (1976). Econometrics. New York: Marcel Dekker, Inc.
- Schmidt, P. and Waud R. (1973). "The Almon Lag Technique and the Monetary and Fiscal Policy Debate." Journal of the American Statistical Association, 68, 11-19.
- Schwarz, G. (1978). "Estimating the Dimension of a Model." The Annals of Statistics, 6, 461-464.
- Seaks, T. G. and Allen, S. D. (1984). "The St. Louis Equation: A Decade Later." Southern Economic Journal, 51, 817-829.
- Shibata, R. (1981). "An Optimal Selection of Regression Variables." Biometrika, 68, 45-54.
- Shorack, G. R. (1982). "Bootstrapping Robust Regression." Communications in Statistics--Theory and Methods, 11, 961-972.

- Simonoff, J. S. and Chih-Ling Tsai. (1986). "Jackknife-Based Estimators and Confidence Regions in Nonlinear Regression." Technometrics, 28, 103-112.
- Sims, C. A. (1980). "Macroeconomics and Reality." Econometrica, 48, 1-48.
- Singh, K. (1981). "On the Asymptotic Accuracy of Efron's Bootstrap." Annals of Statistics, 9, 1187-1195.
- Snee, R. D. and Marquardt, D. W. (1984). "Collinearity Diagnostics Depend on the Domain of Prediction, the Model and the Data." American Statistician, 38, 83-87.
- Sobol, I. M. (1974). The Monte Carlo Method. Chicago: University of Chicago Press.
- Spitzer, J. J. (1982a). "A Primer on Box-Cox Estimation." Review of Economics and Statistics, 64, 307-313.
- Spitzer, J. J. (1982b). "A Fast and Efficient Algorithm for the Estimation of Parameters in Models with Box-and-Cox Transformation." Journal of the American Statistical Association, 73, 760-766.
- Stein, C. M. (1962). "Confidence Sets for the Mean of a Multivariate Normal Distribution." Journal of the Royal Statistical Society, B, 24, 256-285.
- Stein, C. M. (1956). "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution." Proceedings of the Third Berkley Symposium on Mathematical Statistics and Probability. Volume 1. Berkley: University of California Press, 197-206.
- Stein, C. M. (1981). "Estimation of the Mean of a Multivariate Normal Distribution." Annals of Statistics, 9, 1135-1151.
- Stephens, M. A. (1974). "EDF Statistics for Goodness-of-Fit and Some Comparisons." Journal of the American Statistical Association, 69, 730-737.
- Strawderman, W. E. (1978). "Minimax Adaptive Generalized Ridge Regression Estimators." Journal of the American Statistical Association, 17, 623-627.
- Swanepoel, J. W. H. (1986). "A Note on Proving That the (Modified) Bootstrap Works." Communications in Statistics, Theory and Methods, 15, 3193-3203.
- Theil, H. (1971). Principles of Econometrics. New York:

John Wiley and Sons, Inc.

- Toyoda, T. and Wallace, T. D. (1975). "Estimation of Variance After a Preliminary Test if Homogeneity and Optimal Levels of Significance for the Pre-Test." Journal of Econometrics, 3, 395-404.
- Trivedi, P. K. (1978). "Estimation of a Distributed Lag Model Under Quadratic Loss." Econometrica, 46, 1181-1192.
- Trivedi, P. K. and Pagan, A. R. (1979). "Polynomial Distributed Lags: A Unified Treatment." Economic Studies Quarterly, 30, 37-49.
- Ullah, A. (1982). "The Approximate Distribution Function of the Stein-Rule Estimator." Economic Letters, 10, 305-308.
- Ullah, A. (1980). "The Exact, Large Sample and Small Disturbance Conditions of Dominance of Biased Estimators in Linear Models." Economic Letters, 6, 339-344.
- Ullah, A. (1974). "On the Sampling Distributions of Improved Estimators for Coefficients in Linear Regression." Journal of Econometrics, 2, 143-150.
- Ullah, A., Carter, R. A. L. and Srivastava, V. K. (1984). "Sampling Distribution of Shrinkage Estimators and their F-Ratios in the Regression Model." Journal of Econometrics, 25, 109-122.
- U.S. Department of Commerce, Bureau of Economic Analysis. (1985). "An Advanced Overview of the Comprehensive Revision of the National Income and Product Accounts." Survey of Current Business, pp. 19-28.
- U.S. Department of Commerce, Bureau of Economic Analysis. (1985). "Revised Estimates of the National Income and Product Accounts of the United States, 1925-1985." Survey of Current Business, pp. 1-19.
- Vinod, H. D. (1984). "Distribution of a Generalized t-Ratio for Biased Estimators." Economic Letters, 14, 43-52.
- Vinod, H. D. and Ullah, A. (1981). Recent Advances in Regression Methods. New York: Marcel Dekker, Inc.
- Wallich, H. C. (1984). "Recent Techniques of Monetary Policy." Federal Reserve Bank of Kansas City Economic Review, 69, 21-30.

- White, H. and MacDonald, G. M. (1980). "Some Large-Sample Tests for Nonnormality in the Linear Regression Model." Jouranl of the American Statistical Association, 75, 16-31.
- Wu, C. F. J. (1986). "Jackknife, Bootstrap, and Other Resampling Methods in Regression Analysis." Annals of Statistics, 14, 1261-1295.
- Zellner, A. (1971). An Introduction to Bayesian Inference in Econometrics. New York: John Wiley and Sons, Inc.
- Zellner, A. Lee, T. H. (1965). "Joint Estimation of Relationships Involving Discrete Random Variables." Econometrica, 33, 382-394.
- Zellner, A. and Vandaele, W. (1975). "Bayes-Stein Estimators for K Means, Regression and Simultaneous Equation Models." In Studies in Bayesian Econometrics and Statistics. Edited by J. E. Feinberg and A. Zellner. New York: North-Holland.

Vita
Lee C. Adkins

940 Stanford #208
P. O. Box 22727
Baton Rouge, LA 70893

Department of Economics
Oklahoma State University
Stillwater, OK 74078

EDUCATION: A.A., Pensacola Junior College, 1978

 B.S. (cum laude), Marketing, 1980
 Florida State University

 M.S., Economics, 1985
 Louisiana State University

 Ph.D., Economics, 1988
 Louisiana State University

AWARDS: Beta Gamma Sigma
 Alumni Federation Fellowship, 1983-1987

DISSERTATION: Stein-Like Estimation and Inference

DISSERTATION
DIRECTOR: R. Carter Hill

RESEARCH FIELDS: Theoretical and Applied Econometrics

PROFESSIONAL
ASSOCIATIONS: Econometric Society
 American Statistical Association
 Omicron Delta Epsilon
 Southern Economic Association

DOCTORAL EXAMINATION AND DISSERTATION REPORT

Candidate: Lee C. Adkins

Major Field: Economics

Title of Dissertation: Stein-Like Estimation and Inference

Approved:

R. Carter Hill

Major Professor and Chairman

F. Glen Hembry

Dean of the Graduate School

EXAMINING COMMITTEE:

D. J. Sisk

W. Douglas McMillin

William J. Moore

Husain Sarkar

Donald C. Huffman

Date of Examination:

8 July 1988