FOUNDATIONS OF
THE PRINCIPAL ELEMENTS REGRESSION STRUCTURE

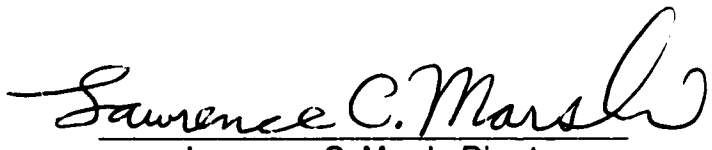
A Dissertation

Submitted to the Graduate School
of the University of Notre Dame
In Partial Fulfillment of the Requirements
of the Degree of

Doctor of Philosophy

by

Kevin Dale Brunson, B.S., M.A.



Lawrence C. Marsh, Director

Department of Economics
Notre Dame, Indiana

April, 1995

FOUNDATIONS OF
THE PRINCIPAL ELEMENTS REGRESSION STRUCTURE

Abstract

by

Kevin Dale Brunson

This dissertation makes three important contributions to econometrics. First, principal elements regression (PER) is shown to be the more general case of such existing procedures as factor analysis, principal components regression (PCR), ordinary least squares (OLS), ridge regression, and some forms of restricted least squares. Second, it introduces three new PER estimators as well as PER versions of existing estimators such as the Mundlak pre-test type and the Stein-rule type. And third, it critiques and refines the correlated notions of the rank of a matrix and the number of linear restrictions on a matrix.

Theoretically, an estimation procedure is derived that offers an improvement over the unbiased least squares and biased principal components estimators when the performance criterion of an estimator is variance reduction subject to orthogonality and dimension constraints. Specifically, with an equal or smaller number of eigenvector values (partial or fractional dimensions) deleted, the trace of the variance-covariance matrix

of the principal elements estimator is shown to be less than or equal to the trace of the variance-covariance matrix of the principal components estimator; it is already widely known that the trace of the variance-covariance matrix of the principal components estimator is less than or equal to the trace of the variance-covariance matrix of the least squares estimator (Fomby, Hill, and Johnson, 1978).

It is also widely known that shrinkage estimators of the Stein-rule type dominate least squares in certain regions of the parameter space under quadratic risk loss functions. This work debuts several principal elements versions of shrinkage estimators and links them to such popular Stein-rule techniques as the positive-part Stein-rule and the limited translation Stein-rule. Further, the PER method is motivated on the same grounds as Stein's limited translation rule.

The performances of the principal elements estimators and their various competitors are evaluated in monte carlo simulations performed on a classical data set and a pseudorandom population using a squared error risk function and an analysis of probability densities.

Dedication

I'll forever be indebted to my wife Billie for urging me to apply to graduate school, for supporting me through the arduous but rewarding years while I matriculated toward my Ph.D., and for not allowing me to quit before the paper work was done. Thus, this document transcends the econometric content--it is a manifestation of her love for me and my love for her...our love for each other

Someday when they are older my daughter Rachel and my son Dane will read this and understand that they had to share a little of me so that I could complete it. Some part of me that could have been theirs is embodied in this document so it belongs to them, too. Looking back, I believe that I balanced the period well because they are both very good kids.

Of course, what would I have done without Larry Marsh! This dissertation literally was a journey, from Washington D.C. to Atlanta to New Orleans to Mexico City, that we traveled together. It probably won't end here--I hope not.

To my other committee members, Paul Kochanowski, Marc Jarsulic, and Scott Maxwell I extend my heartfelt appreciation for their patience and advice. I benefited from their diversity.

Finally, a special thank you to John Peck cf Indiana University South Bend. Because of him I acquired the vision that a thirty-something man with a family could complete this task.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Selecting the 'best' statistical model to solve a prediction or explanation problem via regression techniques involves choosing from a host of approaches even if the researcher has previously framed the problem in the correct theoretical context, therein narrowing the possible choices. For instance, a multiple regression problem where multicollinearity exists is the status quo for economic data and not the exception. Foregoing the issue of quantifying the degree of multicollinearity and, thus, its seriousness, a first decision might be whether or not to use a biased estimator or an unbiased estimator. If the choice is an unbiased estimator like least squares then the correct variables for inclusion still have to be determined--a nontrivial problem since a least squares model with additional irrelevant variables will still be unbiased. And in the presence of multicollinearity the experimenter must also be aware of the shortcoming of unreliable individual parameter estimates[1]. Further, even if least squares is chosen, if data is scarce there remains the complex problem of what data to use to calibrate the model (in-sample) and what data to use to make

[1]Whether the goal is the "best" explanatory model (perhaps for policy planning) or the "best" forecasting model, the reliability of the model decreases as the sample estimates depart from the true plane of collinearity.

predictions (out-sample). Fomby and Hill (1989) demonstrated the uncertain effect on prediction error of differing collinearity structures in the two data subsets. Obviously, using the same data to both derive least squares parameter estimates and then to make predictions generates selection bias (see Miller, 1990). The term selection bias will be explained further in Chapter 5.

Ignoring epistemological concerns about the biased/unbiased schism, a researcher who opts for a biased estimator such as principal components would then be confronted with myriad decisions. A short list would include decisions about the number of components to replace with zeroes[2], whether to employ shrinkage (such as Stein-rule) versions of principal components, and the appropriate loss function. There is a plethora of ad hoc rules on zeroing components; notable among those are pre-test rules employing F-tests and zeroing components where the corresponding eigenvalue is less than one (see Jackson (1991) for a comprehensive treatment of the most popular ones). Similarly, there is a menu of dozens of Stein-rule and other shrinkage estimators from which the researcher can choose as well as many loss functions besides quadratic forms.

Returning to the point I am trying to make, a point which is by no means profound or mystical, the analyst should be overwhelmed by the numerous regression techniques and encumbrances even if the model being considered is the result of cogent reasoning! Suppose an estimator

---

[2]Throughout this work the action of replacing an element of a matrix or vector with a zero will be termed "zeroing." The commonplace term "deleting", used when an entire component is dropped, prevents adoption of a fractional reduction strategy when individual elements of the component vector are not removed.

proposes to replace some of the elements of the eigenvector matrix with zeroes. Completeness necessitates allowing for all possible combinations of elements. That is, having decided upon some number to replace, perhaps by forward inclusion or backward deletion, the technique should consider reinstating some before definitely deciding to zero more. In principal components the number of components is given by the number of explanatory variables, so if X is an $N \times K$ matrix of explanatory variables then there are K candidates for deletion. Letting r represent the number of retained components, then the number of possible combinations is given by

$$\sum_{r=1}^{K} C_r^K = \sum_{r=1}^{K} \frac{K!}{(K-r)!(r)!} = 2^K - 1,$$

subtracting one because the case of zeroing all of the elements is not viable[3]. In the principal elements case, there are K^2 eigenvector elements to select from because each principal component is composed of K principal elements. So the choice becomes unwieldy, but not impossible, for values of K as it approaches 10. When K = 10 there are 1023 principal components combinations and 1.2677 x 10^{30} principal elements combinations. Particularly over the past two decades, insightful minds have conjured a cascade of biased and unbiased estimators that need to be corralled. The economics profession, where 10 explanatory variables is often a simple case, and any others that employ multivariate regression can

---

[3] Note that the instance that all elements are retained is the ordinary least squares result. Hence, OLS is a special case of PER.

benefit from a unifying structure such as principal elements. Indeed, the principal elements structure provides a context to consolidate them.

Even though multicollinearity is only one undesirable characteristic of sample data, it attracts significant amounts of research because its affects are so difficult to mitigate. Among attempts to measure it are the condition number, variance inflation factors, and the rank of a matrix; still, there is no single measurement method which is unarguable.

However, it is known that multicollinearity can inflate variances so researchers focus on the manifestation of multicollinearity as embodied in loss functions such as quadratic risk. This dissertation takes that same turn and uses minimum mean squared error as one basis for estimator comparison. So, here multicollinearity serves as a point of departure and not a feature.

The principal elements framework in this work advances econometric analysis by introducing new statistical models, altering the way information is utilized, and suggesting better ways to judge the performance of competing estimators.

The remainder of this dissertation will unfold as follows. A review of the literature focusing principally on the developments of the past two decades is in Chapter 2. Chapter 3 contains the main body of theoretical material that explains the principal elements structure, analyzes Mundlak-type rules, and explores a few versions of shrinkage estimators. Next, Chapter 4 presents the subject in an easy to visualize pictorial fashion. Chapter 5 continues by explaining the conduct of the monte carlo simulations before Chapter 6 reports the results of those experiments.

Finally, Chapter 7 concludes the dissertation and suggests topics for future research.

CHAPTER 2

LITERATURE REVIEW

Although there are some overlaps, the pertinent literature can simply be divided into that which primarily involves principal components regression and that which primarily involves shrinkage rules.

Fortunately, the principal components literature has been summarized recently by both Jackson (1990) and Oksanen (1988). According to Oksanen, principal components can be divided into three strands--(1) data-descriptive, (2) estimation of the classical linear model, and (3) factor analysis; the first two are significant for my purposes. Jackson's book is a "how-to" manual that tends to emphasize the first strand and also serves as a guide on the evolution of principal components techniques but not theory. Since, as pointed out by Jackson, the literature on principal components is fragmented yet profuse in the past two decades, I will rely on his efforts to pull together this diverse activity.

Principal components has its origins with Karl Pearson in 1901 although today's form was first put forth by Hotelling in 1933. Expansion and adoption accelerated when more powerful computers were introduced in the 1970's and subsequently authors such as Massy (1965), Chen (1974), Greenberg (1975), Fomby, Hill and Johnson (1978), Mundlak (1981), Lee (1986), and Hill and Judge (1987) made contributions to the theory of

principal components. Within the family of principal components estimation rules Greenberg showed that zeroing components corresponding to the smallest eigenvalues brought the greatest reduction in the trace of the variance-covariance matrix of the estimated regression coefficients. Fomby, Hill and Johnson extended this optimality property to the entire class of restricted least squares estimators in the context of considering all possible linear restrictions on the regression coefficients. Simply put, these techniques substituted zeroes for selected values.

Lee's dissertation was a breakthrough in the evolution of principal components because it turned attention to the partial adjustment of each component, a technique he dubbed fractional principal components. He claimed that the fractional structure contained "all the plausible biased estimators useful in combatting (sic) multicollinearity." A reasonable corollary is to investigate the impact of a partial adjustment of the principal elements comprising each component. To date no one has examined the contribution to the variance reduction effort that can be made by alteration of the principal elements, an approach contained herein which admits Lee's family of estimators into the principal elements fold.

A technique which suggests weighting a matrix in the interval [0,1] might be connected to the group of shrinkage estimators such as the popular Stein-rule type first described by James and Stein (1961). In fact, Lee developed a Stein-rule estimator within the context of his fractional principal components. But there is an abundance of writings on the topic of Stein-rules, notable among them are Baranachik (1964), Efron and Morris (1973), Judge and Bock (1978), Stein (1981), Dey and Berger (1983), and Hill and

Judge (1987). Judge and Bock (1978) consolidated existing literature up to the time of their publication and thoroughly examined a number of Stein-rules and other shrinkage algorithms. More recently Judge and Yancey (1986) reported improved rules for estimating linear statistical models that included pretest types and Stein-rule types.

CHAPTER 3

THEORY

3.1 The Principal Elements Structure

Consider the standard regression model

$$y = X\beta + \varepsilon$$

where y is a $t \times 1$ vector of dependent variable values, X is a $t \times k$ matrix of explanatory variable values, β is a $k \times 1$ vector of regression coefficients, and ε is a $t \times 1$ vector of independent disturbances with a common mean of zero and a common variance σ^2.

Note: Because of simplicity, in this section the theory is explained in terms of the extreme case when eigenvector elements are replaced with zeroes. Section 3.3 will explain that zeroing is merely a special case of shrinking elements until they are zero.

3.1.1 Principal Components Regression

Define $Z = XA$ where A is the matrix whose columns a_j are the eigenvectors of the $X'X$ matrix with the columns of A ordered to correspond

to the ordered eigenvalues λ_j where $\lambda_1 > \lambda_2 > \ldots > \lambda_K$. The ordered eigenvalues λ_j also serve as the diagonal elements in the $K \times K$ matrix Λ. The ordinary least squares estimator of β is

[3.1.1] $$\beta^o = (X'X)^{-1} X'y = A(Z'Z)^{-1} Z'y$$

By appropriately partitioning the Z and A matrices as $Z = [Z_1 : Z_2]$ and $A = [A_1 : A_2]$ the principal components regression estimator may be expressed as

[3.1.2] $$\beta^c = A_1 \left(A_1'X'XA_1 \right)^{-1} A_1'X'y = A_1 \left(Z_1'Z_1 \right)^{-1} Z_1'y$$

when A_2 is deleted or

[3.1.3] $$\beta^c = A^c \left(Z'Z \right)^{-1} Z'y \ ,$$

defining $A^c = [A_1 : 0]$ where A_2 is replaced with a matrix of zeros. Of course, since the principal components are orthogonal to one another, the least squares regression run on the reduced set of principal components Z_1 is equivalent to the regression run on the full set of principal components Z with the unwanted results zeroed afterwards using the A^c matrix. These results are equivalent to applying least squares to the model

$$y = X\beta + \varepsilon = ZA'\beta + \varepsilon$$

to minimize

$$\varepsilon'\varepsilon = y'y - 2\beta'AZ'y + \beta'\Lambda\beta$$

subject to the restriction $(A - A^c)'\beta = 0$.

The singular value and spectral value decomposition for the variance-covariance matrix of the least squares estimated regression coefficients are given by

$$Cov(\beta^o) = \sigma^2 (X'X)^{-1} = \sigma^2 A\Lambda^{-1}A' = \sigma^2 \sum_{j=1}^{K} \frac{1}{\lambda_j} a_j a_j' \quad ,$$

where

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & \lambda_j \end{bmatrix} .$$

In terms of the partitioned A matrix, the least squares covariance matrix may be expressed as

[3.1.3] $$Cov(\beta^o) = \sigma^2 A_1\Lambda_1^{-1}A_1' + \sigma^2 A_2\Lambda_2^{-1}A_2'$$

where Λ_1 and Λ_2 are defined in terms of the corresponding partitioning of the Λ matrix

$$\Lambda = \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda 2 \end{bmatrix} .$$

In contrast, the principal components covariance matrix can be written as

[3.1.4] $$Cov(\beta^c) = \sigma^2 A^c \Lambda^{-1} A^{c\prime} = \sigma^2 A_1 \Lambda_1^{-1} A_1' .$$

3.1.2 Principal Elements Regression

If the ultimate objective is to define an estimator that has a smaller mean squared error (MSE) than either least squares or principal components, first direct attention to the variance portion of MSE because that is the part that can more effectively be expressed and controlled. Hence, interest is in a covariance matrix that has a smaller trace than the trace of the β^o covariance matrix given as

$$tr\left[Cov\left(\beta^o\right)\right] = \sigma^2 \sum_{i=1}^{k} \sum_{j=1}^{k} \frac{1}{\lambda_j} a_{ij}^2 = \sigma^2 \left[vec(A)\right]' \left[\Lambda^{-1} \otimes I\right] \left[vec(A)\right]$$

where I is a $k \times k$ identity matrix. This trace may be thought of as the sum of k^2 terms $\left(a_{ij}^2/\lambda_j\right)$ which may be ordered from largest to smallest. PER is defined by selecting the largest of these terms, then the next largest, and so forth, and setting the appropriate a_{ij} values in the original A matrix equal to zero to form the A^e matrix. An ordering (or ranking) operator ^{o} is defined

that creates a $k \times k$ matrix $\overset{o}{A}$ whose elements are integer values reflecting this ordering of individual elements[4].

The principal elements estimator becomes

[3.1.5] $$\beta^e = A^e (Z'Z)^{-1} Z'y = A^e A' \beta^o .$$

This is equivalent to applying least squares to the model

$$y = X\beta + \varepsilon = ZA'\beta + \varepsilon$$

subject to the restriction $\left(A - A^e\right)' \beta = 0$.

3.1.3 Comparing the Covariance Traces

Theorem 1:

1. The trace of the variance-covariance matrix of the principal elements estimator is less than or equal to the trace of the variance-covariance matrix of the principal components estimator with an equal number of eigenvector elements zeroed.

2. The trace of the variance-covariance matrix of the principal elements estimator is less than or equal to the trace of the variance-covariance matrix of the least squares estimator.

---

[4]See Appendix A.1 for an example of a principal elements matrix and an ordering matrix $\overset{o}{A}$

Proof:

The columns of the A matrix are already ordered according to the size of the corresponding eigenvalue (from the largest λ_1 to the smallest λ_k). Designate the first r columns of A as A_1 (columns to be retained) and the last d columns of A as A_2 (columns to be zeroed). Define $\Delta = [\Lambda \otimes I]$, dimension $k \times k$, and specify

$$\Delta = \begin{bmatrix} \Delta_1 & 0 \\ 0 & \Delta_2 \end{bmatrix}$$

where Δ_1 is the $rk \times rk$ diagonal matrix whose diagonal elements repeat each of the r largest eigenvalues k times, and Δ_2 is the $dk \times dk$ diagonal matrix whose diagonal elements repeat each of the d smallest eigenvalues k times (note: $k = r + d$). The trace of $Cov(\beta^o)$ may then be expressed as

$$tr\left[Cov(\beta^o)\right] = \sigma^2 \left[vec(A_1)\right]' \left[\Delta_1^{-1}\right]\left[vec(A_1)\right] + \sigma^2 \left[vec(A_2)\right]' \left[\Delta_2^{-1}\right]\left[vec(A_2)\right] .$$

The principal components estimator β^c has as the trace of its covariance matrix

[3.1.6] $$tr\left[Cov(\beta^c)\right] = \sigma^2 \left[vec(A_1)\right]' \left[\Delta_1^{-1}\right]\left[vec(A_1)\right] .$$

Now, recall the ordering operator ^{o} that specifies the order of the individual elements within the A matrix. This newly ordered matrix $\overset{o}{A}$ is transformed into a vector with the elements of $vec\left(\overset{o}{A}\right)$ ordered to put the diagonal elements of

$$\left[\Lambda^{-1} \otimes I\right]\left[vec(A)\right]\left[vec(A)\right]'$$

in order from largest to smallest down the diagonal. This results in

$$\left[\overset{o}{\Delta}^{-1} \otimes I\right]\left[vec\left(\overset{o}{A}\right)\right]\left[vec\left(\overset{o}{A}\right)\right]'$$

where $\overset{o}{\Delta}$ is the $k^2 \times k^2$ diagonal Δ matrix with the diagonal elements of Δ reordered to correspond to the new ordering of the a_{ij} elements in the new $\overset{o}{A}$ matrix. The trace of the OLS covariance matrix may be expressed in terms of this new ordering as

[3.1.7] $$tr\left[Cov\left(\beta^{o}\right)\right] = \sigma^2\left[vec\left(\overset{o}{A}\right)\right]'\left[\overset{o}{\Delta}^{-1}\right]\left[vec\left(\overset{o}{A}\right)\right].$$

Next, partition the $\overset{o}{A}$ matrix and, correspondingly, the Δ matrix as

$$\overset{o}{A} = \left[\overset{o}{A}_1 : \overset{o}{A}_2\right] \text{ and } \overset{o}{\Delta} = \begin{bmatrix} \overset{o}{\Delta}_1 & 0 \\ 0 & \overset{o}{\Delta}_2 \end{bmatrix}$$

Under this new partitioning we get

$$tr\left[Cov\left(\beta^{o}\right)\right] = \sigma^{2}\left[vec\left(\overset{o}{A}_{1}\right)\right]'\left[\overset{o}{\Delta}_{1}^{-1}\right]\left[vec\left(\overset{o}{A}_{1}\right)\right] + \sigma^{2}\left[vec\left(\overset{o}{A}_{2}\right)\right]'\left[\overset{o}{\Delta}_{2}^{-1}\right]\left[vec\left(\overset{o}{A}_{2}\right)\right].$$

The trace of the covariance matrix for the principal elements estimator may then be expressed as

[3.1.8] $$tr\left[Cov\left(\beta^{e}\right)\right] = \sigma^{2}\left[vec\left(\overset{o}{A}_{1}\right)\right]'\left[\overset{o}{\Delta}_{1}^{-1}\right]\left[vec\left(\overset{o}{A}_{1}\right)\right].$$

Considering the relative size of the individual (a_{ij}^{2}/λ_{j}) elements, this covariance matrix clearly produces a trace smaller or equal to that provided by principal components regression when zeroing the same number of individual eigenvector elements. Therefore, the following inequalities follow directly:

[3.1.9] $$\sigma^{2}\left[vec\left(\overset{o}{A}_{1}\right)\right]'\left[\overset{o}{\Delta}_{1}^{-1}\right]\left[vec\left(\overset{o}{A}_{1}\right)\right] \leq \sigma^{2}\left[vec\left(A_{1}\right)\right]'\left[\Delta_{1}^{-1}\right]\left[vec\left(A_{1}\right)\right]$$

[3.1.10] $$\sigma^{2}\left[vec\left(\overset{o}{A}_{2}\right)\right]'\left[\overset{o}{\Delta}_{2}^{-1}\right]\left[vec\left(\overset{o}{A}_{2}\right)\right] \geq$$

$$\sigma^2 \left[vec(A_2) \right]' \left[\Delta_2^{-1} \right] \left[vec(A_2) \right].$$

In terms of the $\overset{o}{A}$ matrix as defined above, such that A^e has the same ordered structure as the A matrix but the elements in A^e corresponding to those in $\overset{o}{A}_2$ have been replaced with zeros,

$$\begin{aligned} tr\left[Cov\left(\beta^e\right) \right] &= \sigma^2 \left[vec\left(A^e\right) \right]' \left[\Delta^{-1} \right] \left[vec\left(A^e\right) \right] \\ &= \sigma^2 \left[vec\left(A_1^e\right) \right]' \left[\Delta_1^{e\,-1} \right] \left[vec\left(A_1^e\right) \right] . \end{aligned}$$

Since the second term on the right-hand side $\left[\Delta_1^{e\,-1} \right]$ is positive semidefinite, when the same number of individual eigenvector elements are zeroed,

$$tr\left[Cov\left(\beta^e\right) \right] \leq tr\left[Cov\left(\beta^c\right) \right] \leq tr\left[Cov\left(\beta^o\right) \right].$$

This completes the proof.

3.1.4 Linear Restrictions

Once again, consider the usual regression model

$$y = X\beta + \varepsilon$$

and its transformation

$$y = XAA'\beta + \varepsilon = Z\theta + \varepsilon .$$

It could be argued that the proper perspective for understanding a regression problem is in terms of the original data and not the transformed data. Certainly this is the case for an applied problem in economics such as a policy matter. For instance, a governing body would want to know the impact of various public expenditures on private productivity (in other words, the coefficients β_i) and not the impact of ephemeral orthogonal variables (in other words, the coefficients θ_j). Accepting that perspective, a single PCR restriction in θ-space that involves k elements is misleading and should be viewed as k impositions in β-space. In other words, a single linear restriction in β-space can be further decomposed into a number of "partial" linear restrictions, each contributing to that single linear restriction. Currently, there is no term to express this notion so useful terminology will be developed below.

In the case of principal components, the usual expression for a restriction R on the vector of explanatory variables β,

$$R\beta = r,$$

in reality is

$$R\beta = 0,$$

since, because of orthogonality in the PCR technique, decision rules which call for deleting columns of the $k \times k$ eigenvector matrix A is equivalent to zeroing the elements of the selected vectors. The matrix R is a $j \times k$ matrix, $j \leq k$, containing non-sample information about relationships among the individual parameters of β and is said to contain j linear equality restrictions. The simplest case of PCR where only one eigenvector is zeroed amounts to imposing one linear equality restriction on the columns of A but k linear equality restrictions on the rows of A!

To see this better, the PCR restriction in Section 3.1.1, defined as

$$\left(A - A^c\right)' \beta = 0,$$

in 3-space would be

$$\left(\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} - \begin{bmatrix} a_{11} & a_{12} & 0 \\ a_{21} & a_{22} & 0 \\ a_{31} & a_{32} & 0 \end{bmatrix}\right)' \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = 0$$

$$\begin{bmatrix} 0 & 0 & a_{13} \\ 0 & 0 & a_{23} \\ 0 & 0 & a_{33} \end{bmatrix}' \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = 0$$

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ a_{13} & a_{23} & a_{33} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = 0$$

$$[\beta_1 a_{13} + \beta_2 a_{23} + \beta_3 a_{33}] = 0 .$$

To understand how the beta vector is affected, recall that $\beta = A\theta$ and expand,

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

$$\begin{bmatrix} a_{11}\theta_1 + a_{12}\theta_2 + a_{13}\theta_3 \\ a_{21}\theta_1 + a_{22}\theta_2 + a_{23}\theta_3 \\ a_{31}\theta_1 + a_{32}\theta_2 + a_{33}\theta_3 \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

so that

$$\begin{aligned} \beta_1 &= a_{11}\theta_1 + a_{12}\theta_2 + a_{13}\theta_3 \\ \beta_2 &= a_{21}\theta_1 + a_{22}\theta_2 + a_{23}\theta_3 \\ \beta_3 &= a_{31}\theta_1 + a_{32}\theta_2 + a_{33}\theta_3 . \end{aligned}$$

Imposing a restriction on the last column of A, that is $a_{13} = a_{23} = a_{33} = 0$, results in

$$\begin{aligned} \beta_1 &= a_{11}\theta_1 + a_{12}\theta_2 \\ \beta_2 &= a_{21}\theta_1 + a_{22}\theta_2 \\ \beta_3 &= a_{31}\theta_1 + a_{32}\theta_2 . \end{aligned}$$

Thus, it is easily seen that the **single** linear restriction in principal components space, or theta space, involves **three** pieces of information from the beta vector!

Next, consider the PER estimator subject to the same restriction

$$R\beta = 0.$$

In this more flexible estimation framework the restriction would be

$$\left(A - A^{e}\right)' \beta = 0,$$

and in 3 -space case might be

$$\left(\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} - \begin{bmatrix} a_{11} & a_{12} & 0 \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & 0 \end{bmatrix}\right)' \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = 0$$

$$\begin{bmatrix} 0 & 0 & a_{13} \\ 0 & 0 & 0 \\ 0 & 0 & a_{33} \end{bmatrix}' \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = 0$$

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ a_{13} & 0 & a_{33} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = 0$$

$$[\beta_1 a_{13} + \beta_3 a_{33}] = 0 \ .$$

In this instance the beta vector is

$$\beta_1 = a_{11}\theta_1 + a_{12}\theta_2$$
$$\beta_2 = a_{21}\theta_1 + a_{22}\theta_2 + a_{23}\theta_3$$
$$\beta_3 = a_{31}\theta_1 + a_{32}\theta_2 .$$

Thus, it is easily seen that the PER rule involves **only two** pieces of information from the beta vector with a **single** linear restriction! Further, in the simplest case of PER where only one element is zeroed, **only one** piece of information is removed (compared to three for PCR)! Hence, the term *restriction*, which embodies the information lost by many traditional regression techniques that reduce the amount of information composing the beta vector, is inadequate for the principal elements framework. In this simple 3×3 example there are 7 different ways to impose a **single** restriction!

3.1.5 Modulation and Subrestrictions

Thinking in terms of linear restrictions is problematic in the PER shrinkage method of MSE reduction because it is more general than zeroing or deleting eigenvector elements; estimation techniques such as ADHOC, INDIV, and MATRIX can be thought of as fine tuning the signal emanating from a data matrix and in that sense modulate the individual eigenvector elements. Therefore, the term *modulation* is introduced to refer to element-wise constraints or partial restrictions that are called *subrestrictions.* Modulating a matrix involves imposing partial restrictions on the columns (or rows) that are called subrestrictions.

3.1.6 Matrix Rank and Matrix Mode

The notion of the rank of a matrix is related to the notion of a restriction. One definition of rank is the minimum of the number of linearly independent columns or rows. For the $k \times k$ eigenvector matrix A with rank k, if PCR dictates that one column should be zeroed, the rank would be reduced to $k - 1$. However, in the PER example above where only two of the elements of the third column were zeroed, the rank of the matrix is still k, a scalar measure of the number of unrestricted columns. Obviously, the rank of a matrix is misleading as to the number of subrestrictions imposed.

To avoid losing this sort of information the term *mode* is defined as the number of un-modulated elements. Thus, unlike PCR, other PER methods can achieve variance reduction by modulating the eigenvector matrix without reducing its rank but will reduce the mode.

This simple illustration highlights the inadequacy of using either the number of restrictions or the rank of a matrix as a comparative descriptive statistic whereas the mode of a matrix does not obfuscate the actual constraints imposed on the beta vector. In this scenario, the PCR eigenvector matrix has a rank of 2 and a mode of 6 while the PER eigenvector matrix has a rank of 3 and a mode of 7.

3.2 Mundlak-type Rules

There is an abundance of criteria for guiding the researcher on selecting the correct model for a particular application. Miller (1990) is a superb source for both a bibliography and a taxonomy of these rules;

among the most popular are Mallows' C_p, Akaike's AIC, and Schwarz' BIC. George and Foster (1990) suggest a newer one based upon a calculation of risk inflation.

The decision on how many elements an estimator should delete involves considering a whole host of different stopping rules. Certainly for any principal components selection stopping rule, a corresponding principal elements selection stopping rule can be devised that is roughly equivalent. For example, under principal components regression one nonstochastic stopping rule is to delete all those components that correspond to eigenvalues less than one where the data has been scaled to a zero mean and unit variance. Of course, this rule implies zeroing components when the reciprocal of an eigenvalue is greater than one. This stopping rule inequality may be written as:

$$\frac{a_{1j}^2}{\lambda_j} + \ldots + \frac{a_{Kj}^2}{\lambda_j} = \frac{1}{\lambda_j} \sum_{i=1}^{K} a_{ij}^2 = \frac{1}{\lambda_j} > 1 .$$

This formulation says that the sum of k terms must be greater than one, or, on average each term must be greater than one divided by k. This suggests the following nonstochastic stopping rule for principal elements regression:

zero all eigenvector elements that correspond to the terms greater than one divided by k.

In other words, form the $\overset{o}{A}$ matrix by setting to zero all a_{ij} values that correspond to the $\left(a_{ij}^2/\lambda_j\right)$ terms that meet the requirement:

$$\frac{a_{ij}^2}{\lambda_j} > \frac{1}{k} .$$

This defines a nonstochastic stopping rule for PER which is defined solely in terms of the independent variables and completely ignores values of the dependent variable. Such rules have the advantage of having well-defined statistical properties and hypothesis testing procedures, but also possess unbounded risk since they ignore the dependent variable values and thus at times can be highly inaccurate.

Alternatively, any number of stochastic stopping rules could be devised. One such rule could be modeled after the principal components stopping rule that says delete all those components that have statistically insignificant t-statistics in the regression run on all of the principal components.

A sophisticated variation of pre-test rules is Mundlak's F-Testing for nonsignificant sets of principal components (Mundlak, 1981) which uses the F statistic

$$F = \frac{1}{D} \sum_{j=1}^{D} t_j^2$$

where D is the number of components zeroed and where

$$t_j^2 = \frac{1}{\hat{\sigma}^2 \left[\frac{a_{1j}^2}{\hat{\theta}_j^2 \lambda_j} + \frac{a_{2j}^2}{\hat{\theta}_j^2 \lambda_j} + \ldots + \frac{a_{kj}^2}{\hat{\theta}_j^2 \lambda_j} \right]} .$$

Adapting Mundlak's rule to the principal elements regression context, the corresponding F statistic is:

[3.2.1] $$F_m = \frac{1}{R\,\hat{\sigma}^2} \sum_{j=1}^{D} \left[\frac{1}{\sum_{i=1}^{d_j} \frac{a_{ij}^2}{\hat{\theta}_j^2\ \lambda_j}} \right]$$

where R is the rank of the $\left(A - A^e\right)$ matrix, D is now redefined as the number of eigenvectors which have at least one zeroed element, d_j is the number of a_{ij} elements zeroed for the j^{th} of the D eigenvectors with at least one element zeroed, and F_m is defined to be the minimum value of all possible such F values with exactly m eigenvector elements zeroed. Under this modified Mundlak rule the optimal value of F_m (and, therefore, the optimal set of $m$eigenvector values to zero) is such that either:

1. $F_m < F^m_{n-R}$ table value and $F_{m+1} > F^{m+1}_{n-R}$ table value.

2. $F_1 > F^1_{n-K}$ table value and the matrix X has rank K.

3. $F_{K-1} < F^{K-1}_{n-K}$ table value and the matrix X has rank 1.

3.3 Shrinkage Estimators

This section relaxes the severe constraint of zeroing individual eigenvector elements so that a partial reduction in magnitude is allowed as well.

3.3.1 Fractional Principal Components Regression

Here are proposed three different, but related, fractional principal elements regression (FPCR) estimators (see Lee and Birch (1986) and Lee (1986)). In this formulation shrinkage is applied to a principal component in its entirety but each component can be weighted differently. Traditional principal components uses a binary weighting scheme whereby the effect of a component is either removed entirely by zeroing (a weight of 0) or it is left unchanged (a weight of 1). FPCR relaxes this restriction for the particular component that is the last candidate for zeroing. Begin with the standard regression model

$$y = X\beta + \varepsilon$$

and restate it as

$$y = XAA'\beta + \varepsilon = Z\theta + \varepsilon$$

where A is the matrix whose columns are the eigenvectors of $X'X$ and Z is the matrix of principal components as defined in section 3.1.1. Next, perform the transformation

$$y = ZF'F\beta + \varepsilon = Z^f\beta^f + \varepsilon$$

where F is a $k \times k$ diagonal matrix subject to $(0 \leq f_{ij} \leq 1)$. Noting that

$$Z^f = ZF' = XAF' ,$$

then FPCR can be expressed in terms of the principal components matrix A as

$$A^f = AF' ,$$

so that a fractional principal components estimator of β is

[3.3.1] $$\beta^f = A^f (Z'Z)^{-1} Z'y$$

when converted to the PER framework.

Lee specified the diagonal elements of F as $f_{ij} = 1$ for the r components retained and $f_{ij} = 0$ for the $k - (r + 1)$ components zeroed. However, the $(r + 1)^{th}$ $f_{ij} = f_s$, which corresponds to the pivotal component as a candidate for inclusion or zeroing, is weighted in the interval [0, 1] so that the component is only partially retained.

In a PER context, this fractional method is a special case when $k - 1$ components are entirely zeroed or entirely retained (as in traditional PCR techniques) and only one is partially removed. In other words, the

principal elements that correspond to f_s are all reduced by the same amount instead of selectively zeroing or reducing some.

The optimal fraction for f_s is chosen according to the rule

[3.3.2] $$f_s = \frac{\theta_j^2}{\theta_j^2 + \frac{\sigma^2}{\lambda_j}}$$

which is the MSE minimizing solution with respect to the f_j 's.

Since θ_j and σ^2 are unknown this optimal estimator is problematic; nevertheless, in this case it is possible that the orthogonality property of the principal components coefficients, θ_j's, might enable the OLS estimates $\hat{\theta}_j$'s to provide robust substitutes that may help retain "near" optimality for the FPCR estimator.

3.3.2 Principal Elements Shrinkage

The first PER method, ADHOC, is a completely ad hoc method originally formulated by analogy with the simple problem of estimating a single population mean. The second method, INDIV, is aimed at reducing the MSE of each of the estimated regression coefficients of the original model individually. Since it adjusts each estimated coefficient separately and ignores cross product terms it is essentially a nonmatrix approach. The third method, MATRIX, is a full matrix method that is developed with a concern for minimizing the trace of the MSE matrix of all of the estimated

regression coefficients simultaneously so it makes adjustments to the cross product terms and not just the matrix diagonal.

ADHOC

The first of these estimation methods, ADHOC, merely defines the elements of the principal elements biasing matrix A^a in a manner analogous to the 'optimal' solution to the problem of finding the mean squared error minimizing estimate of a single sample mean:

$$\tilde{\mu} = \frac{\mu^2}{\mu^2 + \frac{\sigma^2}{n}} \bar{x}.$$

The analogous form ADHOC shrinks each principal element individually,

$$a_{ij}^a = \frac{\theta_j^2}{\theta_j^2 + \frac{\sigma^2}{\lambda_j}} a_{ij}$$

and estimates β as

[3.3.3] $$\beta^a = A^a (Z'Z)^{-1} Z'y .$$

Note the similar form to Lee's FPCR, 3.3.2, a consequence of deriving both estimators as MSE minimizing solutions. The ADHOC form of FPCR is

$$a_j^f = f_s a_j = \frac{\theta_j^2}{\theta_j^2 + \frac{\sigma^2}{\lambda_j}} a_j \ .$$

Since θ_j and σ^2 are unknown this optimal estimator is problematic; nevertheless, in this case it is possible that the orthogonality property of the principal components coefficients, θ_j's, might enable the OLS estimates θ_j^o's to provide robust substitutes that may help retain "near" optimality for the ADHOC estimator.

INDIV

The second principal elements regression method, INDIV, is derived to minimize the MSE of the individual estimated regression coefficients as follows:

$$\begin{aligned}
MSE\left(\beta_i^e\right) &= \text{var}\ \beta_i^e + \left(E\beta_i^e - \beta_i\right)^2 \\
&= \text{var}\ \beta_i^e + \left(E\beta_i^e - E\beta_i^o\right)^2 \\
&= \text{var}\ \beta_i^e + E\left(\beta_i^e - \beta_i^o\right)^2 - \text{var}\left(\beta_i^e - \beta_i^o\right) \\
&= \text{var}\ \beta_i^e + E\left(\beta_i^e - \beta_i^o\right)^2 - \text{var}\ \beta_i^e - \text{var}\ \beta_i^o + 2\,Cov\left(\beta_i^e, \beta_i^o\right) \\
&= E\left(\beta_i^e - \beta_i^o\right)^2 - \text{var}\ \beta_i^o + 2\,Cov\left(\beta_i^e, \beta_i^o\right).
\end{aligned}$$

Next, note the following expressions for β_i^o and β_i^e which are written as linear functions of the ordinary least squares estimators θ_j^o for $j = 1, \ldots, k$ of the principal components regression coefficients, the

eigenvector values a_{ij} of the original $X'X$ matrix, and the unknown principal elements a_{ij}^e which are the population constants to be estimated:

$$\beta_i^o = \sum_{j=1}^{K} a_{ij}\theta_j^o \qquad \text{and} \qquad \beta_i^e = \sum_{j=1}^{K} a_{ij}^e\theta_j^o \ .$$

Substituting these expressions yields the following formulation:

$$MSE\left(\beta_i^e\right) = E\left[\sum_{j=1}^{K}\left(a_{ij}^e - a_{ij}\right)\theta_j^o\right]^2 - \text{var}\sum_{j=1}^{K} a_{ij}\theta_j^o + 2\,Cov\left[\sum_{j=1}^{K} a_{ij}^e\theta_j^o\ ,\ \sum_{j=1}^{K} a_{ij}\theta_j^o\right]$$

$$MSE\left(\beta_i^e\right) = E\left[\sum_{j=1}^{K}\left(a_{ij}^e - a_{ij}\right)\theta_j^o\right]^2 - \text{var}\sum_{j=1}^{K} a_{ij}\theta_j^o$$

$$+2\,E\left[\left\{\sum_{j=1}^{K} a_{ij}^e\left(\theta_j^o - E\theta_j^o\right)\right\}\left\{\sum_{j=1}^{K} a_{ij}\left(\theta_j^o - E\theta_j^o\right)\right\}\right]$$

But by the very nature of principal components regression, the θ_j^o uncorrelated with one another, i.e.,

$$E\left[\left(\theta_j^o - E\theta_j^o\right)\left(\theta_k^o - E\theta_k^o\right)\right] = 0$$

for $j \neq k$, so the MSE expression becomes,

$$MSE\left(\beta_i^e\right) = E\left[\sum_{j=1}^{K}\left(a_{ij}^e - a_{ij}\right)\theta_j^o\right]^2$$

$$-\sum_{j=1}^{K} a_{ij}^2 \text{ var } \theta_j^o + 2\sum_{j=1}^{K} a_{ij}^{e2} \text{ var } \theta_j^o .$$

expanding the first term while noting that $E\left(\theta_j^o\theta_k^o\right) = E\theta_j^o E\theta_k^o$,

$$MSE\left(\beta_i^e\right) = \sum_{j=1}^{K}\left(a_{ij}^e - a_{ij}\right)^2 E\theta_j^2$$

$$+2\sum_{k>j}\left(a_{ij}^e - a_{ij}\right)\left(a_{ik}^e - a_{ik}\right)E\theta_j^o E\theta_k^o$$

$$-\sum_{j=1}^{K} a_{ij}^2 \text{ var } \theta_j^o + 2\sum_{j=1}^{K} a_{ij}^e a_{ij} \text{ var } \theta_j^o$$

Now, from the definition of variance

$$E\left(\theta_j^{o2}\right) = \text{var } \theta_j^o + \left(E\theta_j^o\right)^2$$

so that

$$MSE\left(\beta_i^e\right) = \sum_{j=1}^{K}\left(a_{ij}^e - a_{ij}\right)^2\left[\text{var } \theta_j^o + \left(E\theta_j^o\right)^2\right]$$

$$+2\sum_{k>j}\left(a_{ij}^e - a_{ij}\right)\left(a_{ik}^e - a_{ik}\right)E\theta_j^o E\theta_k^o$$

$$-\sum_{j=1}^{K} a_{ij}^2 \text{ var } \theta_j^o + 2\sum_{j=1}^{K} a_{ij}^e a_{ij} \text{ var } \theta_j^o .$$

and since under principal components regression θ_j^o and θ_k^o are unbiased estimators of θ_j and θ_k respectively, and since var $\theta_j^o = \frac{\sigma^2}{\lambda_j}$, we get

$$MSE\left(\beta_i^e\right) = \sum_{j=1}^{K}\left(a_{ij}^e - a_{ij}\right)^2\left[\frac{\sigma^2}{\lambda_j} + \theta_j^2\right]$$

$$+2\sum_{k>j}\left(a_{ij}^e - a_{ij}\right)\left(a_{ik}^e - a_{ik}\right)\theta_j\theta_k$$

$$-\sum_{j=1}^{K} a_{ij}^2 \frac{\sigma^2}{\lambda_j} + 2\sum_{j=1}^{K} a_{ij}^e a_{ij} \frac{\sigma^2}{\lambda_j} .$$

Expanding, combining, and canceling terms results in

$$MSE\left(\beta^e\right) = \sum_{j=1}^{K} a_{ij}^{e\,2} \frac{\sigma^2}{\lambda_j} + \left[\sum_{j=1}^{K}\left(a_{ij}^e - a_{ij}\right)\theta_j\right]^2 .$$

Taking the first derivative for a minimum or maximum results in

[3.3.4] $$\frac{dMSE\left(\beta_i^e\right)}{da_{ij}^e} = 2a_{ij}^e \frac{\sigma^2}{\lambda_j} + 2\theta_j \sum_{j=1}^{K} \left(a_{ij}^e - a_{ij}\right)\theta_j = 0 .$$

Checking that the second derivative is positive for a minimum,

$$\frac{d^2MSE\left(\beta_i^e\right)}{d^2a_{ij}^e} = 2\frac{\sigma^2}{\lambda_j} + 2\theta_j^2 > 0 .$$

Following from the first derivative expression

$$a_{ij}^e \frac{\sigma^2}{\lambda_j} + \theta_j \sum_{j=1}^{K} a_{ij}^e\theta_j - \theta_j \sum_{j=1}^{K} a_{ij}\theta_j = 0 \quad .$$

Now multiply through by $\lambda_j\theta_j$ to get

$$\sigma^2 a_{ij}^e\theta_j + \lambda_j\theta_j^2 \sum_{j=1}^{K} a_{ij}^e\theta_j = \lambda_j\theta_j^2 \sum_{j=1}^{K} a_{ij}\theta_j ,$$

then sum over $j = 1, \ldots, K$ to get

$$\sigma^2 \sum_{j=1}^{K} a_{ij}^e\theta_j + \sum_{j=1}^{K} \lambda_j\theta_j^2 \sum_{j=1}^{K} a_{ij}^e\theta_j = \sum_{j=1}^{K} \lambda_j\theta_j^2 \sum_{j=1}^{K} a_{ij}\theta_j$$

or,

$$\left[\sigma^2 + \sum_{j=1}^{K} \lambda_j \theta_j^2 \right] \sum_{j=1}^{K} a_{ij}^e \theta_j = \sum_{j=1}^{K} \lambda_j \theta_j^2 \sum_{j=1}^{K} a_{ij} \theta_j ,$$

and finally

$$\sum_{j=1}^{K} a_{ij}^e \theta_j = \frac{\sum_{j=1}^{K} \lambda_j \theta_j^2 \sum_{j=1}^{K} a_{ij} \theta_j}{\left[\sigma^2 + \sum_{j=1}^{K} \lambda_j \theta_j^2 \right]} .$$

Returning to the derivative expression

$$a_{ij}^e \frac{\sigma^2}{\lambda_j} + \theta_j \sum_{j=1}^{K} a_{ij}^e \theta_j - \theta_j \sum_{j=1}^{K} a_{ij} \theta_j = 0 ,$$

dividing all terms by θ_j yields

$$a_{ij}^e \frac{\sigma^2}{\lambda_j \theta_j} = \sum_{j=1}^{K} a_{ij} \theta_j - \sum_{j=1}^{K} a_{ij}^e \theta_j .$$

Now substitute for $\sum_{j=1}^{K} a_{ij}^e \theta_j$ from above and rearrange to estimate individual a_{ij}

[3.3.5]
$$a_{ij}^e = \frac{\lambda_j \theta_j \sum_{j=1}^{K} a_{ij} \theta_j}{\left[\sum_{j=1}^{K} \lambda_j \theta_j^2 + \sigma^2 \right]} .$$

Estimates of β are then calculated as

[3.3.6] $$\beta^i = A^i (Z'Z)^{-1} Z'y$$

where $A^i = \left\{ a_{ij}^e \right\}$; that is, it is made up of individual a_{ij}^e that are the typical ij^{th} element of the A^i matrix.

Since θ_j and σ^2 are unknown this optimal estimator is problematic; nevertheless, in this case it is possible that the orthogonality property of the principal components coefficients, θ_j's, might enable the OLS estimates θ_j^o's to provide robust substitutes that may help retain "near" optimality for the INDIV estimator.

MATRIX

The third PER method MATRIX is derived from the matrix expression for the mean squared error of prediction as follows:

$$\begin{aligned} PMSE(\beta^e) &= E\left[\left(y^e - Ey \right)' \left(y^e - Ey \right) \right] \\ &= E\left[\left(X\beta^e - X\beta \right)' \left(X\beta^e - X\beta \right) \right] \\ &= E\left[\left(\beta^e - \beta \right)' X'X \left(\beta^e - \beta \right) \right] \end{aligned}$$

and since a scalar is equal to its trace

$$PMSE\left(\beta^e\right) = E\left\{tr\left[X'X\left(\beta^e - \beta\right)\left(\beta^e - \beta\right)'\right]\right\}$$

$$= E\left\{tr\left[X'X\left(A^e\theta^o - A\theta\right)\left(\theta^{o\prime}A^{e\prime} - \theta'A'\right)'\right]\right\}$$

$$= tr\left[X'X\,E\left(A^e\theta^o\theta^{o\prime}A^{e\prime} - A^e\theta^o\theta'A' - A\theta\theta^{o\prime}A^{e\prime} + A\theta\theta'A'\right)\right]$$

$$= tr\left[X'X\left(A^e\left(E\theta^o\theta^{o\prime}\right)A^{e\prime} - A^e\left(E\theta^o\right)\theta'A' - A\theta\left(E\theta^{o\prime}\right)A^{e\prime} + A\theta\theta'A'\right)\right.$$

Note: Since the principal components estimator θ^o is an unbiased estimator of the population principal components coefficient θ, we have $E\theta^o = \theta$. Also, by definition of the covariance of θ^o:

$$Cov\theta^o = E\left(\theta^o - E\theta^o\right)\left(\theta^o - E\theta^o\right)'$$

$$= E\theta^o\theta^{o\prime} - E\theta^o\left(E\theta^o\right)' = \sigma^2\Lambda^{-1}$$

Of course, this implies

$$E\theta^o\theta^{o\prime} = \sigma^2\Lambda^{-1} + \theta\theta'$$

which, when substituted into the above expression for $PMSE\left(\beta^e\right)$ results in

$$PMSE\left(\beta^e\right) = tr\Big[X'X\Big(\sigma^2 A^e \Lambda^{-1} A^{e'} + A^e \theta\theta' A^{e'} - A^e \theta\theta' A' - A\theta\theta' A^{e'} + A\theta\theta' A' \Big)\Big]$$

$$= tr\Big[\sigma^2 X'XA^e \Lambda^{-1} A^{e'} + X'XA^e \theta\theta' A^e \theta\theta' - X'XA^e \theta\theta' A' - X'XA\theta\theta' A^{e'} + X'XA\theta\theta' A' \Big]$$

$$= tr\Big[\sigma^2 A^{e'} X'XA^e \Lambda^{-1} + A^{e'} X'XAA^e \theta\theta' - X'XA^e \theta\theta' A' - X'XA\theta\theta' A^{e'} + X'XA\theta\theta' A' \Big],$$

and the final expression is

[3.3.7] $$PMSE\left(\beta^e\right) = \sigma^2 tr\Big[A^{e'} X'XA^e \Lambda^{-1} \Big] + tr\Big[A^{e'} X'XAA^e \theta\theta' \Big] - tr\Big[X'XA^e \theta\theta' A' \Big] - tr\Big[X'XA\theta\theta' A^{e'} \Big] + tr\big[X'XA\theta\theta' A' \big].$$

Next, to minimize this predicted mean squared error, employ the following three rules from the calculus of matrices:

(1) $$\frac{d\, tr(WZ')}{d\, Z} = W,$$

(2) $$\frac{d\, tr(VZW)}{d\, Z} = V'W',$$

$$(3)\quad \frac{d\,tr(Z'VZW)}{d\,Z} = V'ZW' + VZW\ .$$

When the resulting first derivative is set equal to zero

$$\frac{d\,PMSE(\beta^e)}{d\,A^e} = 2\sigma^2 X'XA^e\Lambda^{-1} + 2\,X'XA^e\theta\theta' - 2\,X'XA\theta\theta'$$
$$= 0.$$

The second derivative is positive definite for a minimum,

$$\frac{d^2\,PMSE(\beta^e)}{d\,A^{e^2}} = 2\sigma^2 X'X\Lambda^{-1} + 2\,X'X\theta\theta' > 0\ .$$

Multiplying the first derivative expression by $\frac{1}{2}(X'X)^{-1}$ and collecting terms yields

$$A^m\left[\theta\theta' + \sigma^2\Lambda^{-1}\right] = A\theta\theta'$$

or,

[3.3.8] $$A^m = A\theta\theta'\left[\theta\theta' + \sigma^2\Lambda^{-1}\right]^{-1}\ .$$

Finally, the MATRIX estimator of β is expressed as

[3.3.9] $$\beta^m = A^m (Z'Z)^{-1} Z'y .$$

Again, as with ADHOC and INDIV, since θ_j and σ^2 are unknown this optimal estimator is problematic; nevertheless, in this case it is possible that the orthogonality property of the principal components coefficients, θ_j's, might enable the OLS estimates θ^o's to provide robust substitutes that may help retain "near" optimality for the MATRIX estimator.

3.3.3 Stein-type Shrinkage

Stein-type shrinkage has frequently produced a smaller MSE than competing estimators but this class of estimators suffers because exact analytic solutions are not available (another limitation is that there must be at least three regressors present). Consequently, their performance must be evaluated in monte carlo simulations which makes general statements about their superiority unsupported outside of the parameter space of the experiment. However, they display a tendency to be minimax; doubtlessly, researchers suspect that they may generally produce a smaller MSE than most traditional estimation techniques so they are considered in this experiment.

PP

In particular, truncated Stein estimators are superior in terms of squared error loss (Judge and Bock, 1978). One useful member of the

family of positive part estimators (PP) proposed by Baranachik (1964) and Stein (1966) and subsequently analyzed by many researchers is

$$\theta_i^+ = \left(1 - \min\left[1 or \frac{c}{\theta^{o\prime}\theta^o}\right]\right)\theta^o ,$$

where $0 \leq c \leq 2(K-2)$.

Of the many subsequent permutations of this rule offered, a variation used in Hill and Judge (1987) that explicitly utilizes an estimator β^q of β, produces

[3.3.10] $$\beta_i^+ = \left[1 - \frac{a^*}{F(\cdot)}\right]\left(\beta - \beta^q\right) + \beta^q ,$$

where $F(\cdot)$ is a random variable with an F distribution having J and $(T-K)$ degrees of freedom. Of course, J is the number of elements that have been zeroed, β^q is a general estimate of the β vector, and a^* has been chosen subject to

$$o \leq a^* \leq \frac{2(T-K)(J-2)}{(T-K+2)J} .$$

This restriction makes PP minimax in terms of MSE.

LTS

Numerous authors have assessed the characteristics and performance of Charles Stein's (1981) unbiased risk estimator both

analytically and through numerical methods. Notable among the analytical examinations are Judge, Hill, and Bock (1990) and important monte carlo investigations are contained in Dey and Berger (1983), Hill and Judge (1987, 1989), and Adkins (1991). Stein's shrinkage estimator, dubbed a limited translation Stein-rule (LTS), was an enhancement of a modified James-Stein estimator originally proposed in Efron and Morris (1971, 1972).

Consider a $K \times 1$ vector γ and its estimate of an individual γ_i

$$[3.3.11] \qquad \gamma_i^{(\rho)} = \left[1 - \frac{(\rho - 2)\sigma^2 \min\left\{ 1 or \omega_{(\rho)} / |\gamma_i| \right\}}{\sum_{i=1}^{K} \left[\min\left\{ \gamma_i^2 or \omega_{(\rho)}^2 \right\} \right]} \right] \gamma_i,$$

where ρ is a large fraction of K and the order statistics corresponding to ρ are $\omega_i = |\gamma_i|$ and $\omega_1 < \omega_2 < \ldots < \omega_k$. This form of a Stein-rule limits shrinkage to only those parameters that are relatively precisely estimated and leaves those that have larger error margins untouched. This is appealing because only those parameters that are providing 'good' information are allowed to improve their signal so that the reduction in variance might offset the small increase in bias.

An operational version of this more sophisticated Stein-rule estimate of an individual β_i is

[3.3.12]

$$\gamma_i^{o(\rho)} = \left[1 - \frac{\left(\rho^* - 2\right) \hat{\sigma}^2 \min\left\{ 1 or \left(\omega_{(\rho)} / \left| \gamma_i^o \right| \right) \right\}}{\sum_{i=1}^{K} \left[\min\left\{ \gamma_i^{o2} or \omega_{(\rho)}^2 \right\} \right]} \right]^+ \gamma_i^o$$

,

and β_i is estimated as $\left(\beta_i^{(\rho)} = \Lambda^{\frac{1}{2}\prime} A \gamma_i^{(\rho)} \right)$.

This form was first given in Dey and Berger (1983) as a positive-part variation--in other words, a truncated estimator. This truncated estimator is made useful by letting the data choose the value for ρ^*, $(\rho \geq 3)$, which maximizes

$$\frac{(\rho - 2)^2}{\sum_{i=1}^{K} \left[\min\left\{ \gamma_i^{o2} or \omega_{(\rho)}^2 \right\} \right]}.$$

The maximum value of this function of ρ is chosen first by a recursive procedure for $(K - 2)$ iterations. Individual $\gamma_i^{(\rho)}$ are then calculated in another recursive procedure utilizing [3.3.12] that ultimately yields an estimator of β.

This LTS is an attractive choice because the numerical studies in both Dey and Berger (1983) and Hill and Judge (1987) suggest that this expression may be minimax. Since their version of [3.3.12] merely operates

in standardized principal components space it is a simple exercise to express it in terms of principal components and principal elements as will be shown in the following two sections.

3.3.4 A Truncated Principal Components Estimator

Again, consider again the usual regression model

$$y = X\beta + \varepsilon$$

as explained in section 3.1. The necessary variant for implementing [3.3.10] is

$$\begin{aligned} y &= X\beta + \varepsilon = XAA'\beta + \varepsilon \\ &= Z\theta + \varepsilon = Z\Lambda^{-\frac{1}{2}}\Lambda^{\frac{1}{2}}\theta + \varepsilon \end{aligned}$$

and in final form,

$$y = W\gamma + \varepsilon \quad .$$

The least squares estimator of γ,

[3.3.13] $$\gamma^o = W'y \sim N\left(\gamma, \theta^2 I_K\right) \quad ,$$

can be decomposed into a principal components form to be

[3.3.14] $$\gamma^o = W'y = \Lambda^{-\frac{1}{2}}Z'y = \Lambda^{-\frac{1}{2}}A'X'y.$$

Now, eschewing the issue of choosing how many and which principal components to zero, for PCR the matrix of eigenvectors will be partitioned into $A^c = [A_1 : A_2] = [A_1 : 0]$ where A_2 has been replaced with a matrix of zeros. The principal components estimator in [3.1.12] is

$$\beta^c = A^c (Z'Z)^{-1} Z'y \quad .$$

Substituting A^c for A in [3.3.12] gives the principal components version of the LTS γ as

[3.3.15] $$\gamma^c = W^{c'} y \sim N(\gamma, \sigma^2 I_K) \quad .$$

Hill and Judge (1987) suggested investigating the consequences of such a hybrid estimator instead of the least squares variety using a Mundlak pre-test estimator to make the decision on how many principal components to zero. As well, the principal elements version also warrants investigation.

3.3.5 A Truncated Principal Elements Estimator

The LTS estimator [3.3.10] can also be expressed as a member of the principal elements family of estimators as follows. Recalling that one particular approach to PER presented before is to select the largest of the terms $\left(a_{ij}^2 / \lambda_j \right)$ then the next largest, and so forth and setting the

corresponding a_{ij} values in the original A matrix equal to zero to form the A^e matrix, the principal elements estimator in [3.1.5] is

$$\beta^e = A^e (Z'Z)^{-1} Z'y.$$

Substituting A^e for A in [3.3.12] gives the principal components version of the LTS γ as

[3.3.16] $$\gamma^e = W^{e'} y \sim N\left(\gamma, \sigma^2 I_K\right) .$$

This γ^e estimate can also be found by substituting in [3.3.14] the principal elements matrix M as defined by

$$M = \Lambda^{-1} A'^2$$

where M is the $k \times k$ matrix

$$M = \begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1k} \\ m_{21} & m_{22} & & m_{2k} \\ \vdots & & \ddots & \vdots \\ m_{k1} & m_{k2} & \cdots & m_{kk} \end{bmatrix},$$

with individual m_{ij} defined as

$$m_{ij} = \frac{a_{ij}^2}{\lambda_j} \text{ , and } (i, j = 1, 2, \ldots, K) ,$$

to yield

[3.3.17] $$\gamma^e = M^{\frac{1}{2}} X'y .$$

3.3.6 Principal Elements and Restricted Regression

When correct model specification is cast in the restricted least squares form

[3.3.18] $$\beta^o = \beta + (X'X)^{-1} R' \left[R(X'X) \right]^{-1} (r - R\beta)$$

with the restriction of the form $R\beta^o = r$ on the OLS estimate β^o of β, nonlinear search algorithms like Gauss-Newton, Newton-Raphson, Marquardt, and DFP are enlisted to solve a minimization or maximization problem. As stated in section 3.1.2, in the principal elements case

$$R = \left(A - A_0\right)'$$

where

$$A_0 = \left(A - A^e\right)$$

so that

$$R = A^{e'} .$$

When $r = 0$ the subtraction operation results in

$$(r - R\beta) = (0 - R\beta) = A^{e'}\beta.$$

Iteratively solving for the minimum value for R would appear to be efficient if the Newton-Raphson method could be utilized, but it is computationally expedient to avoid the calculation of second derivatives. Therefore, methods such as Gauss-Newton, DFP, and Marquardt, among others, are appropriate since they do not require a second derivative. Further, the choice of algorithm is data dependent as each one performs better in some settings than in others (Greene, 370).

A Gauss-Newton algorithm truncated after the linear portion works by accepting starting values for R and calculates a new value for R at each iteration. When the difference between R_{i+1} and R_i is equal at a chosen criterion, convergence is declared and an extremum for R has been approximated. However, since the shape of the function is unknown there is insufficient knowledge about the second-order conditions; that is, it is uncertain if it is a maximum or minimum, and if it is a minimum, is it local or global?

The algebraic expression of the Gauss-Newton algorithm is given in terms of the following function Y,

$$Y = f(R) + \varepsilon ,$$

subject to

$$f(R) = (X'X)^{-1} R' \left[R(X'X)^{-1} R' \right]^{-1} RA,$$

where Y is a $K \times K$ matrix, R is a $J \times K$ matrix of a subset of the eigenvector matrix, $f(R)$ is a $K \times K$ matrix, and ε is a $K \times K$ matrix of disturbances with a common mean of 0 and a common variance of σ^2.

The first order Taylor series expansion of $f(R)$ is

$$f(R) \approx f(R_0) + f^d(R_0)(R - R_0)$$

(f^d denotes the first derivative)

so that

$$Y = f(R_0) + f^d(R_0)(R - R_0) + \varepsilon^*.$$

Define

$$Y_0^* \equiv Y - f(R_0)$$

so that

$$Y_0^* = f^d(R_0)(R - R_0) + \varepsilon^*$$

or

[3.3.19] $$Y_0^* = f^d(R_0)R - f^d(R_0)R_0 + \varepsilon^*.$$

Then, define

$$Y_0^{**} \equiv Y_0^* - f^d(R_0)R_0$$

and it follows that

[3.3.20] $$Y_0^{**} \equiv f^d(R_0)R_0 + \varepsilon^*.$$

The least squares estimate of R is

$$R_1^o = \left\{\left[f^d(R_0)\right]'\left[f^d(R_0)\right]\right\}^{-1}\left[f^d(R_0)\right]' Y_0^{**}$$

or more generally

$$R_{i+1}^o = \left\{\left[f^d(R_i)\right]'\left[f^d(R_i)\right]\right\}^{-1}\left[f^d(R_i)\right]' Y_i^{**}.$$

Alternatively

$$R_1^o = R_0 + \left\{\left[f^d(R_0)\right]'\left[f^d(R_0)\right]\right\}^{-1}\left[f^d(R_0)\right]' Y_0^*$$

or more generally

$$[3.3.21] \qquad R_{i+1}^{o} = R_i + \left\{ \left[f^d(R_i) \right]' \left[f^d(R_i) \right] \right\}^{-1} \left[f^d(R_i) \right]' Y_i^*$$

Now, finding $\left[f^d(R_0) \right]$ is not straightforward in practice because solving for it requires using the calculus of matrices. If the matrices were vectors the solutions would be relatively simple, but when they are not simple vectors the dimensions of the normal derivatives explode on the order of j and k. So, the following rules are introduced:

$$\left[f^d(R) \right] = \frac{\partial f(R)}{\partial R} = \frac{\partial \left[(X'X)^{-1} R' \left\{ R(X'X)^{-1} R' \right\}^{-1} RA \right]}{\partial R}$$

$$= \left[(X'X)^{-1} \otimes I_j \right] I_{j,k} \left[\left(R(X'X)^{-1} R' \right)^{-1} RA \otimes I_k \right]$$

$$-\left[(X'X)^{-1} \otimes I_j \right] \left[\left(R(X'X)^{-1} R' \right)^{-1} \otimes I_j \right]$$
$$vec\left(I_j' \right) vec' \left[(X'X)^{-1} R' \right] \left[\left(R(X'X)^{-1} R' \right)^{-1} \otimes I_k \right] \left[RA \otimes I_k \right]$$

$$-\left[(X'X)^{-1} \otimes I_j \right] \left[\left(R(X'X)^{-1} R' \right)^{-1} \otimes I_j \right] \left[R(X'X)^{-1} \otimes I_j \right]$$
$$I_{j,k} \left[\left(R(X'X)^{-1} R' \right)^{-1} \otimes I_k \right] \left[RA \otimes I_k \right]$$

$$+ vec\left[\left(R(X'X)^{-1} R' \right)^{-1} R(X'X)^{-1} \right] vec'(A) \ .$$

An advantage of this formulation is that R^0 can now be respecified in vector terms as

$$[3.3.22]\quad vec\left(R_{i+1}^0\right) = \left\{\left[f^d(R_i)\right]'\left[f^d(R_i)\right]\right\}^{-1}\left[f^d(R_i)\right]' vec\left(Y^{**}\right).$$

This is the correct form for programming the least squares solution of the Gauss-Newton algorithm and is the basis for the more efficient (in this instance) Marquardt method which takes the form

$$[3.3.23]\quad vec\left(R_{i+1}^0\right) = \left\{\left[f^d(R_i)\right]'\left[f^d(R_i)\right] + \lambda\ diag\left[f^d(R_i)\right]'\left[f^d(R_i)\right]\right\}^{-1} \left[f^d(R_i)\right]' vec\left(Y^{**}\right).$$

The additional term $\lambda\ diag\left[f^d(R_i)\right]'\left[f^d(R_i)\right]$ merely changes the step length of the adjustment between iterations and lambda takes on an initial value like 10^{-3} and may be adjusted by a factor such as 10 depending upon differences between iterations.

As observed by Greene (370-372), there is a dearth of criteria regarding the selection of a method of nonlinear optimization and the correct choice is a function of the particular data being studied.

3.4 Estimator Set

To summarize, the following estimators of β were presented (β^q is a general estimator of β), where A is the matrix whose columns are the orthonormal eigenvectors of $X'X$ so that $Z = XA$ (the matrix of principal components) and $\theta = A'\beta$.

[3.1.1] OLS $\beta^o = A(Z'Z)^{-1}Z'y$

Ordinary least squares with the complete A matrix.

[3.1.2] PCM $\beta^c = A^c(Z'Z)^{-1}Z'y$

Partition A to yield $A^c = [A_1 : 0]$ according to a Mundlak pre-test type rule.

[3.1.5] PEM $\beta^e = A^e(Z'Z)^{-1}Z'y$

Redefine A as A^e, a specially ordered matrix also calculated according to a Mundlak pre-test type rule.

[3.3.1] FPCR $\beta^f = A^f(Z'Z)^{-1}Z'y$

A^f is chosen according to Lee's fractional estimation method.

[3.3.3] ADHOC $\beta^a = A^a(Z'Z)^{-1}Z'y$

A^a is a form of A analogous to the solution for the mean squared minimizing estimate of a single sample mean.

[3.3.6] INDIV $\beta^i = A^i(Z'Z)^{-1}Z'y$

A^i is the principal elements specification that shrinks each of the a_{ij} individually.

[3.3.9] MATRIX $\beta^m = A^m(Z'Z)^{-1}Z'y$

A^m is the matrix version of INDIV that shrinks the elements concurrently.

CHAPTER 4

INTUITIVE INTERPRETATION

4.1 Unbiasedness

Although the objective is to get estimates as close to the true population values as possible, for mathematical convenience researchers often limit their search for quadratic risk (mean squared error) minimizing estimators to the class of unbiased estimators. This is unfortunate, for although it is true that analytical solutions to the quadratic risk minimization problem are often hard to come by, those who limit themselves by gamely searching the class of unbiased estimators may well be forfeiting the contest.

Consider the usual specification of risk as the expected value of the loss function $L\left(\beta, \beta^q\right)$

$$\rho\left(L\left(\beta, \beta^q\right)\right) = E\left(L\left(\beta, \beta^q\right)\right).$$

If the first criterion is to minimize the risk ρ, it becomes a problem to find an estimator that is admissible. For instance, it is well known that in this context even Stein-type estimators perform best in only some regions of the parameter space and a generalization as to the necessary conditions has

not been forthcoming.[5] A researcher is left with no clear-cut rule on how to choose the one technique that dominates for a particular problem being studied.

Thus, a second condition is often added, such as unbiasedness, invariance, or sufficiency. However, the imposition of the second rule should not render the first one unobtainable and with the unbiased restriction that is precisely what occurs! This is due to the existence of the Cramer-Rao lower bound. On one hand, the Cramer-Rao lower bound is an encouragement to those searching within the class of unbiased estimators because it provides a lower bound for the variance of all unbiased estimators of the specified population parameter being estimated. On the other hand, the Cramer-Rao lower bound establishes a region of variance reduction which unbiased estimators may never enter. Fortunately, biased estimators that can and often do have variances that fall below, and potentially well below the Cramer-Rao lower bound, are able to enter this region of improved variance reduction.

To see this key point more clearly consider the following dart board diagram.

---

[5]The original Stein estimators are superior to OLS/MLE in all regions of the parameter space.

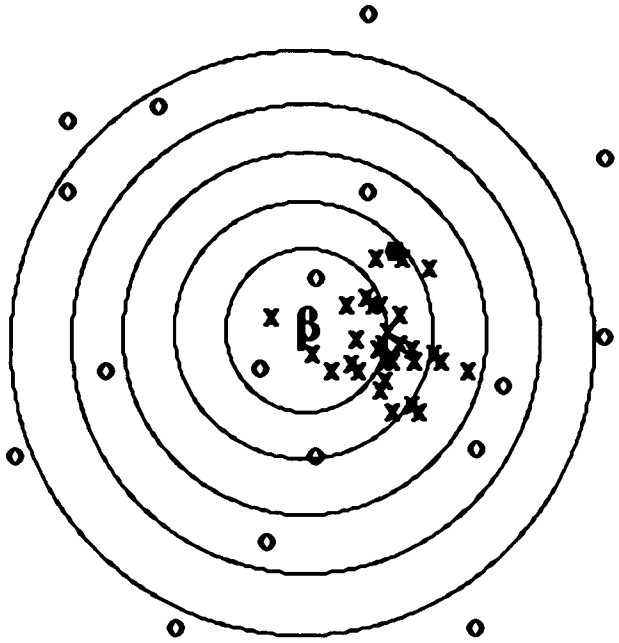


FIGURE 4.1

UNBIASED "O" ESTIMATOR AND BIASED "X" ESTIMATOR

Since the center of the target is the true population parameter β, by definition the unbiased "**o**" estimator (such as OLS) β^o has an expected value $E(\beta^o) = \beta$ so the distribution of β^o is centered at β. Furthermore, if β^o is an efficient estimator of β as determined by the Cramer-Rao lower bound, then β^o has the smallest possible variance of all estimators of β in the class of unbiased estimators. The problem is that this smallest possible variance may be quite large as is generally true in cases of high multicollinearity. Consequently, it is quite possible to find a biased

estimator β^x represented by the "x" values that has a smaller mean squared error than any unbiased estimator.

As is evident in Figure 4.1, the biased "x" estimator may well generate values much closer to the true population value β than the unbiased "o" estimator, and it is entirely possible that some of those estimates will be exact. Of course, if the direction of the bias was known it could be "corrected" to make it unbiased, but often the direction of the bias is unknown. Also, the Cramer-Rao lower bound suggests that it may be impossible to produce an unbiased estimator with as low a variance and, perhaps, as low a mean squared error as is possible with biased estimators. Bear in mind that each "o" and "x" value represents an entire sample of values used to estimate it. For example, a single "o" value may represent 6,000 families as in the Panel Study of Income Dynamics. An investigator only gets one estimate for a given sample. Moreover, we do not know for a given sample where that single "o" value or "x" value will fall relative to the target's center. All we know is that in some sense the "x" values are "more likely" to be close to the target center than the "o" values so there is less risk with the biased estimator.

4.2 Matrix Interpretation

The following hypothetical matrix examples illustrate the action of the various estimators in terms of the reparameterized coefficient vector θ^o where

$$\theta^o = (Z'Z)^{-1} Z'y \text{ and } Z = XA$$

and clearly illustrate that all others are a special case of principal elements regression[6].

- The ordinary least squares estimator, $\beta^o = A\theta^o$, uses all of the elements of the θ^o vector. Thus, the eigenvector matrix would be:

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}$$

- A restricted least squares estimator which simply deletes two of the original explanatory variables would have the form $\beta^r = A^r\theta^o$ and the eigenvector matrix would be:

$$A^r = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

- A PCR estimator which drops the last two principal components would have the form $\beta^c = A^c\theta^o$ where:

---

[6]Appendix A.2 lists some of the actual matrices from the simulations.

$$A^c = \begin{bmatrix} a_{11} & a_{12} & 0 & 0 \\ a_{21} & a_{22} & 0 & 0 \\ a_{31} & a_{32} & 0 & 0 \\ a_{41} & a_{42} & 0 & 0 \end{bmatrix}$$

- A regression based on a factor analysis with two factors, where the first factor represents three original explanatory variables and the second factor represents the second, would have the form $\beta^{fa} = A^{fa}\theta^o$ and produce an eigenvector matrix:

$$A^{fa} = \begin{bmatrix} a_{11} & 0 & 0 & 0 \\ 0 & a_{12} & 0 & 0 \\ a_{31} & 0 & 0 & 0 \\ a_{41} & 0 & 0 & 0 \end{bmatrix}$$

- A FPCR estimator such as Lee's which zeroes the last principal component and reduces the third by f_s, where $(0 \leq f_s \leq 1)$, would have the form $\beta^f = A^f\theta^o$. The eigenvector matrix would appear as:

$$A^f = \begin{bmatrix} a_{11} & a_{12} & f_s a_{13} & 0 \\ a_{21} & a_{22} & f_s a_{23} & 0 \\ a_{31} & a_{32} & f_s a_{33} & 0 \\ a_{41} & a_{42} & f_s a_{43} & 0 \end{bmatrix}$$

- The more flexible PER estimator which weights each of the a_{ij} individually, $f_{ij}^a = f_{ij}a_{ij}$ subject to $(0 \leq f_{ij} \leq 1)$, would have the form $\beta^e = A^e\theta^o$:

$$A^e = \begin{bmatrix} f_{11}^a & f_{12}^a & f_{13}^a & f_{14}^a \\ f_{21}^a & f_{22}^a & f_{23}^a & f_{24}^a \\ f_{31}^a & f_{32}^a & f_{33}^a & f_{34}^a \\ f_{41}^a & f_{42}^a & f_{43}^a & f_{44}^a \end{bmatrix}$$

Whereas PCR uses the same subset of the θ^o vector (that is, the corresponding elements of the same columns are used) in determining each element of the β^c vector, PER can use all of the elements of the θ^o vector in determining the composition of the β^e vector. Thus, each of the elements of the β^e vector can be based on a different subset of the θ^o vector and the dimension of that subset can be different as well.

To see this better, by comparison in the above examples,

$$\beta_1^c = \left(a_{11}\theta_1^o\right) + \left(a_{12}\theta_2^o\right)$$

whereas

$$\beta_1^e = \left(f_{11}^a\theta_1^o\right) + \left(f_{12}^a\theta_2^o\right) + \left(f_{13}^a\theta_3^o\right) + \left(f_{14}^a\theta_4^o\right)$$

and

$$\beta_4^c = \left(a_{41}\theta_1^o\right) + \left(a_{42}\theta_2^o\right)$$

whereas

$$\beta_4^e = \left(f_{41}^a\theta_1^o\right) + \left(f_{42}^a\theta_2^o\right) + \left(f_{43}^a\theta_3^o\right) + \left(f_{44}^a\theta_4^o\right).$$

Since, in PER, the weighting scheme need not zero an element nor even shrink it at all, β_i^e could be equal to β_i^o if it is providing unambiguous information (that is, it is not involved in the multicollinearity thread that connects some of the variables).

In the case where elements are zeroed, it is obvious that PER, by being free to search for the combination of smallest eigenvector elements, has the potential to reduce variance without increasing bias by a greater amount than PCR when the same number or fewer eigenvector elements are zeroed! In the more general case of PER shrinkage, it has the potential to reduce variance without increasing bias by a greater amount than PCR without zeroing eigenvector elements!

Going further, as long as PER shrinkage does not introduce linearities into the eigenvector matrix, there is the potential for a smaller MSE than PCR without reducing its rank. Similarly, there is the potential for a smaller variance than OLS without reducing the rank of the matrix.

4.3 Geometric Interpretation

Another way to gain an understanding of the PER structure is to look at a graphical example of the impact of zeroing in two dimensions.

The original linear statistical model $y = X\beta + \varepsilon$ with least squares estimator $\beta^o = (X'X)^{-1}X'y$ maps the T rows of the X matrix into a K-dimensional space. In the case where $K = 2$, Figure 4.2 locates vector $x_{(1)} = (x_{11}, x_{12})$ onto axes b_1 and b_2.

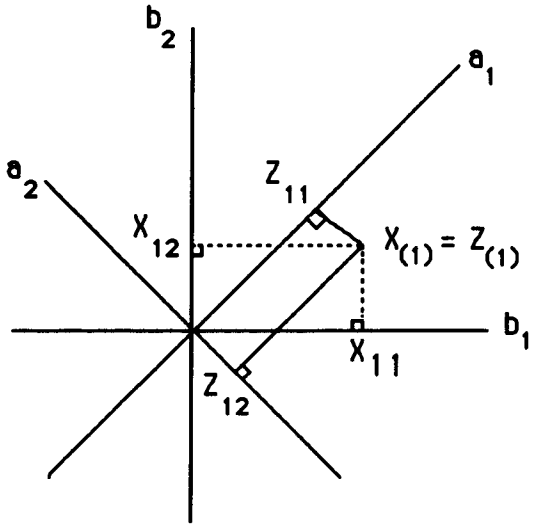


FIGURE 4.2
RELATIONSHIP OF ORIGINAL X VARIABLES
AND PRINCIPAL COMPONENTS, Z

The same point located by the X's and b's can be reexpressed in terms of the Z's and a's by using the eigenvector matrix A to transform the X variables into their principal component variables Z with the relation $Z = XA$. The value of this operation is that the transformation into principal components organizes the data scatter (data ellipse) along its major and minor axes as shown in Figure 4.3.

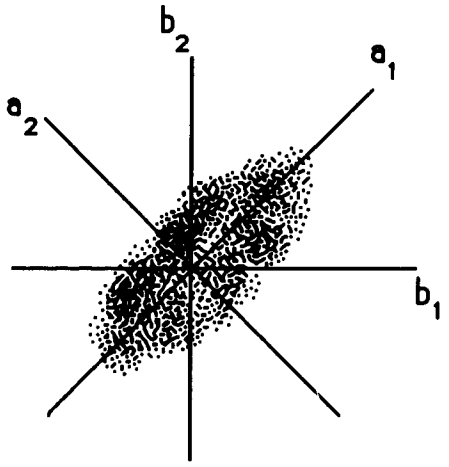


FIGURE 4.3
ALIGNMENT OF DATA ELLIPSE

The depiction also reveals the relationship among variances, eigenvalues and eigenvectors. By construction the first principal component Z_1 captures as much of the variation in the X variables as any one linear combination of the X variables could possibly do. Thus the variance of Z_1 which is its eigenvalue λ_1 represents the data scatter along the a_1 axis which represents the direction defined by the first eigenvector. In turn, the second principal component Z_2 captures as much of the remaining variation in the X's as any second linear combination that is orthogonal to the first linear combination could possibly do. The variance of Z_2 is its eigenvalue λ_2 which measures the dispersion of the data scatter along the axis whose direction is defined by the eigenvector a_2. Clearly, from Figure 4.3 and 4.4a, $\lambda_1 > \lambda_2$. The greater variation along the a_1 axis implies that the information provided by the data along the a_2 axis is less precise than that along the a_1 axis.

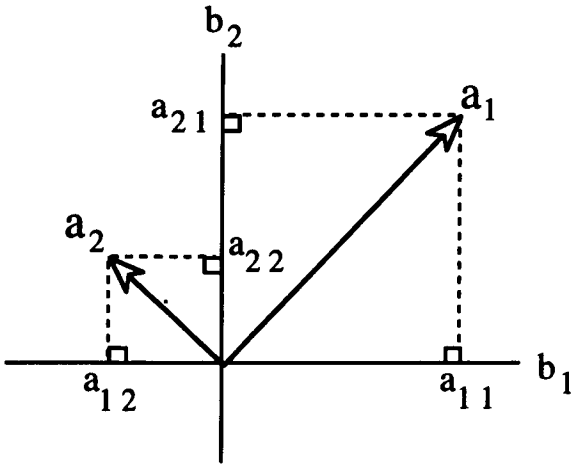


FIGURE 4.4a
ALIGNMENT OF DATA ALONG EIGENVECTORS

Additional components are generated in an analogous manner where each additional principal component is orthogonal to each of the preceding principal components with variance λ_j such that $\lambda_{j-1} > \lambda_j > \lambda_{j+1}$. So, what happens when a component is deleted? Suppose that λ_2 is zero, or close to zero, so that deleting a_2 would be the appropriate restriction. As seen in Figure 4.4b, eliminating a_2, the direction with the least information, would collapse the data ellipse onto a_1, the direction that contains the most information.

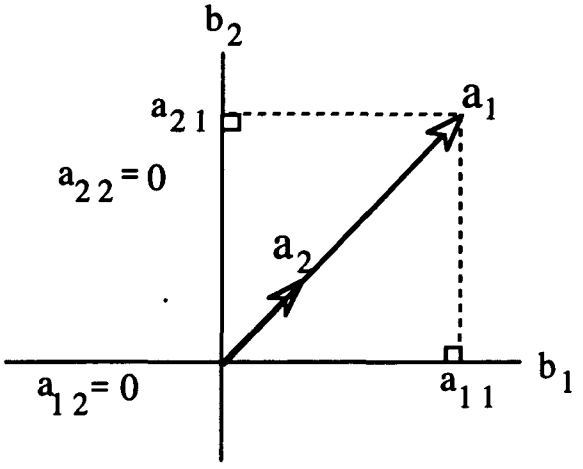


FIGURE 4.4b
TOTAL COLLAPSE OF a_2 VECTOR IN PCR

The implication is that a_2 represents an "extra" dimension that is unstable and provides very little information (even though it may contain an amount of useful information). Therefore, estimation using only a_1 is more stable, reliable, and uses almost all of the information in the original data.

Principal elements deletion is analogous in that it represents a mapping of eigenvectors in the original space spanned by b_1 and b_2 as shown in Figure 4.4c.

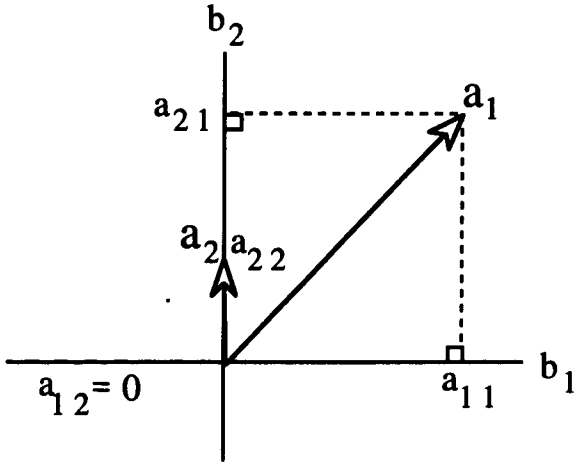


FIGURE 4.4c
PARTIAL COLLAPSE OF a_2 VECTOR IN PER

Instead of collapsing the a_2 vector onto the a_1 vector completely, principal elements regression in this case only partially collapses the a_2 vector by setting $a_{12} = 0$ and, thereby, only partially forcing the a_2 vector to conform to the a_1 direction. This implies less confidence in the information suggested by a_{12} relative to that offered by a_{22}, a_{21}, or a_{11} by some measure. That measure might be the size of the a_{ij} values themselves, or, more likely, some transformation of them such as comparing the $\left(a_{ij}^2/\lambda_j\right)$ values to minimize variance for a given number of elements. If minimizing quadratic loss is the objective, then using information from the stochastic dependent variable Y_i would be preferable to just sticking with nonstochastic information such as a_{ij} and λ_j which come solely from the nonstochastic design matrix, X. One advantage of using the Y_i information,

other than the obvious fact that it's central to the problem, is that only a rule that uses the Y_i information can maneuver to avoid unbounded risk (i.e. it is well known that deletion rules based only on the X information will always have unbounded risk in some region of the parameter space). A reasonable testing strategy for deciding which of the $a_{ij} = 0$ restrictions to impose would be to impose those that seem most compatible with the observed data. This strategy is consistent with Mundlak's F-Testing for nonsignificant restrictions, various Stein-type methods for variable selection, as well as Bayesian posterior odds ratio selection procedures.

CHAPTER 5

SIMULATIONS[7]

The possibilities for realizing great gains by using the principal elements structure is revealed by the results of monte carlo experiments performed first on a designed pseudorandom population and subsequently on the classic Gorman and Toman dataset.

5.1 Selection Bias

In Chapter 1 the term selection bias as used in Miller (1990) was mentioned. He made the distinction between omission bias resulting from leaving variables out of the regression, and selection bias which results from using the same data to both select the model and predict the dependent variable.

To understand omission bias, begin with the standard regression model

$$y = X\beta + \varepsilon$$

[7]All computer calculations were conducted in SAS (release 6.08, TS 415) PROC/IML on the University of Notre Dame's IBM 9121 mainframe running under VM 1.1.

and divide the explanatory variables X into two subsets. Suppose that the model selected from the variables retained in subset A of the data (B subset remains) is

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

The expected value of β_A is

$$\begin{aligned}
E(\beta_A) &= (X_A'X_A)^{-1}X_A'y \\
&= (X_A'X_A)^{-1}X_A'X\beta \\
&= (X_A'X_A)^{-1}X_A'(X_A \vdots X_B)\beta \\
&= (X_A'X_A)^{-1}(X_A'X_A \vdots X_A'X_B)\beta \\
&= (X_A'X_A)^{-1}(X_A'X_A\beta_A + X_A'X_B\beta_B) \\
&= (X_A'X_A)^{-1}(X_A'X_A\beta_A) + (X_A'X_A)^{-1}(X_A'X_B\beta_B) \\
&= \beta_A + (X_A'X_A)^{-1}X_A'X_B\beta_B
\end{aligned}$$

Thus, the second term on the right side, called omission bias, is the bias that accrues from omitting the B subset from the regression. Monte carlo techniques make it possible to avoid this source of bias.

Selection bias is defined as the difference between the expected value of β_A conditional upon A having been selected and the unconditional expected value of β_A:

$$\text{Selection bias} = E(\beta_A \mid \text{subset A chosen}) - E(\beta_A).$$

The second term on the right is the amount of bias regardless of which subset is chosen and the first term represents the instance where the conditions necessary for the selection of β_A are present. In a more exhaustive search procedure both would hold while only the second term would hold when a less intensive search method is used. That is, bias increases as the search for the best estimator becomes more involved. Recalling also from Chapter 1, the new PER estimators have K^2 candidates for treatment while the more traditional estimators have only K. Therefore, it is possible that the Mundlak version of PER will generate greater selection bias than PCR and OLS yet not as much as an exhaustive search method. On the other hand, the new PER forms ADHOC, INDIV, and MATRIX do not search the principal elements matrix but use the available information to produce an estimate in a single calculation and thereby are not guilty of inflating the selection bias.

5.2 Loss Function

The objective in estimating a regression coefficient or any other population parameter is to obtain an estimate that is as close as possible to the true population value; in other words, to minimize the risk function

$$\rho\left(\beta, \beta^q\right) = E\left(L\left(\beta, \beta^q\right)\right).$$

The cost of missing the true population parameter β by estimating its value to be β^q instead of β, may sometimes be expressed as the error, $\left(\beta^q - \beta\right)$, with a quadratic loss function:

[5.2.1] $$L = \left(\beta^q - \beta\right)^2$$

and a corresponding risk function:

$$\rho = E(L) = E\left(\beta^q - \beta\right)^2$$

where the risk is thus defined as the expected value of the corresponding loss function. The quadratic risk function is also known as the mean squared (or square) error which numerous authors have routinely shown to be exactly equal to the variance plus the bias squared

$$MSE\left(\beta^q, \beta\right) = \operatorname{cov}\left(\beta^q\right) + bias\left(\beta^q\right)^2 .$$

Alternative expressions of the risk function besides mean squared error are mean absolute value, mean absolute cubic error, and mean quadratic error . They differ in that as the power of the function increases more weight is given to extreme-valued data points (the outliers). The selection of the best risk function is up to the researcher, but mean squared error is the most prevalent so it is the criterion employed here.

Using MSE implies that biased estimators are being employed because this criterion accounts for the amount of bias generated by the

estimator. Given a measure of risk, such as MSE, attention can be turned to the problem of how to compare the various alternatives. In other words, how does one set the rules so that the playing field is level for all methods? The monte carlo simulations described below will demonstrate that the principal elements structure is a viable context for examining the way in which competing estimators rank their respective parts.

5.3 The Traditional Regression Model[8]

The traditional linear regression model

$$y = X\beta + \varepsilon$$

is based on the following assumptions regarding the error term,

- Normally distributed errors ε_i,
- Errors have a mean of zero $E(\varepsilon_i) = 0$,
- Homoskedastic errors $Var(\varepsilon_i) = \sigma^2$,
- Nonautocorrelated errors $Cov(\varepsilon_i, \varepsilon_j) = 0 \; for (i \neq j)$,

and this assumption about X,

- Nonstochastic explanatory variables.

The traditional regression model assumes a nonstochastic X for convenience even though it is more likely that economic data is random rather than fixed. For instance, frequently only a single sample is obtained

---

[8]See Kmenta (1986), chapters 7 and 8 for a more thorough discussion.

rather than repeated samples. Kmenta (1986) pointed out that this was only a superficial dichotomy because even randomly generated values can be deemed as fixed since the samples are drawn from a finite population.

Thus, the stochastic property is not crucial to an evaluation of an estimator, but on the other hand the relationship between X and ε is. There are 3 possible dependencies,

- X and ε are independent
- X and ε are contemporaneously uncorrelated
- X and ε are not independent or contemporaneously uncorrelated.

The first is the most important for least squares in order to retain its properties of unbiasedness and efficiency and asymptotic properties of efficiency, consistency, and normality. It is even sufficient that only the weaker second condition hold for least squares to retain its asymptotic properties. However, when the third condition does not hold least squares loses its desirable properties.

The experiment on the Gorman and Toman data set is consistent with the classical assumptions whereas a feature of the pseudorandom data set is the stochastic nature of the explanatory variable matrix X. Two of the experiments conducted on it, one where X and ε are independent and a second where X and ε are dependent, examine the independence/dependence assumption.

5.4 Datasets

A decision was made to subject the estimator set to extreme conditions that are often encountered empirically. The variances of the Gorman and Toman data were made arbitrarily large via specification of the coefficient of variation (explained below); in this way the inflated variances that often accompany severe multicollinearity were modeled.

Similarly, the variance of the pseudorandom data was affected by generating a relatively large error vector. In addition, simulations with this data reflect the small sample properties of the estimator set because the sample size was only ten. It was also possible to manipulate the dependency between the explanatory variables and the error vector via the random number generator seed. Using different seeds insured independence where the least squares estimates are unbiased (Gauss Markov Theorem). In contrast, using the same seed induced dependence between the explanatory variables and the error vector, in which case least squares results are biased.

Appendices A.3 and A.4, respectively, pertain to the Gorman and Toman data and pseudorandom data.

5.4.1 Classic: Gorman and Toman

The Gorman and Toman dataset was the basis for the study by Hill and Judge (1987) that is part of the lineage of this work. It comprises 36 observations on a manufacturing process--ten predictor variables were used by Hill and Judge, as well as here, because it provided a benchmark for the

results of the experiments. Even though the true characteristics of a population cannot be known in practice, since they are assumed to be known in simulations, the performance of the principal elements structure relative to OLS, PCR, or other regression techniques can be evaluated based upon this dataset or other empirically obtained datasets.

The following methodology was employed. Preliminary OLS estimates of the beta vector of regression coefficients was first obtained from the unscaled design matrix (Table 5.1). The "true" coefficients BETA were then arbitrarily assigned values so that some were close to the OLS results, some were not, and some had sign reversals. In this way the OLS vector and the population vector were somewhat oriented in the same direction, a method which gave OLS an advantage.

TABLE 5.1
GORMAN AND TOMAN PRELIMINARY-OLS
AND POPULATION COEFFICIENTS

method	β1	β2	β3	β4	β5	β6	β7	β8	β9	β10
OLS	5	-8	-287	-24	-1530	283	8	18	0.27	-0.26
BETA	4	-6	-296	-15	-1300	350	-1	25	-0.28	0.26

Next, an error vector was generated from SAS's NORMAL function calculated as

$$E = \text{NORMAL(SEED)} * \text{SIGMA}$$

where SEED = 16132280649 and SIGMA was calculated as

$$\text{SIGMA} = \text{CV} * \text{YBAR} .$$

The coefficient of variation CV is a unitless number normally derived as the ratio of the estimated standard deviation and the mean of the Y vector. Since the true population standard deviation is unknown, and since it is measured in the units of the dependent variable making it practically unbounded, deriving it by specifying CV avoids making guesses about its magnitude, thus making this measure invariant with respect to choice of units of measurement. CV values up to 2.0 are observed and that was the value chosen (Brunson and Marsh (1991d) explored a range of CV values from .25 up to 2.0). It should be noted that the common practice is to assume $\sigma = 1$ for computational convenience and not because it is any more likely than any other value. The sample coefficient of variation was .28.

A draw from the population of possible dependent variable values was then simulated from the usual model

$$y = X\beta + \varepsilon$$

where ε = "E", β = "BETA", and X, the original design matrix, is fixed in repeated samples. This was repeated N = 5000 times and summary statistics on MSE and estimates of β for each estimator compiled where the MSE at each iteration was calculated as

---

[9]The same value used by Hill and Judge (1987). They chose sigma = 1.

$$MSE^q = \left(\beta^q - \beta\right)' \left(\beta^q - \beta\right)$$

and subsequently averaged over the 5000 experiments.

Whereas the small dimensions of the pseudorandom dataset precluded using Stein-type shrinkage rules, this dataset has ten so additional estimators were added to the set. The treatment PP is the positive-part Stein-type and the treatment DB is the empirical form of LTS. Further, extending the suggestion of Hill and Judge (1987) to condition the γ-vector with a Mundlak pre-test estimator, these experiments conditioned it with PCM, PEM, ADHOC, INDIV, and MATRIX as well as OLS. Thus, there are a total of 18 estimators to compare: OLS, PCM, PEM, ADHOC, INDIV, and MATRIX plus PP and DB versions of each.

The presence of multicollinearity in data can be precisely identified by the existence of at least one zero eigenvalue. In practice data is seldom that cooperative so it is left to the researcher to decide what small value is equivalent to zero. Another popular method to measure multicollinearity is to calculate the square root of the ratio of the largest to smallest eigenvalue, or, the condition number

$$\kappa = \left(\frac{\lambda_1}{\lambda_k}\right)^{\frac{1}{2}},$$

where the eigenvalues are obtained from the fully standardized correlation matrix; otherwise, data measured in different units will affect the condition

number. For the Gorman and Toman data the condition number of the correlation matrix is 6.88484 / .02744 = 250.91 while the condition number of the product matrix is 528,363 / .007 = 75,480,428. There is no precise limit above which a condition number is a manifestation of multicollinearity, but anything above 30 is often accepted as a strong indicator of its presence.

A second, contrasting scenario to high multicollinearity is no linear dependencies; that is, equal eigenvalues. This was accomplished by fixing each eigenvalue according to

$$\lambda_i^* = \frac{\sum_{i=1}^{k} \lambda_i}{k} .$$

TABLE 5.2
EIGENVALUES OF THE GORMAN AND TOMAN DATA

Product Matrix (X'X) Eigenvalues									
λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}
528363	32899	951	362	162	98	5.80	1.33	.16	.007
Correlation Matrix Eigenvalues									
λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}
6.88484	0.90926	0.64067	0.47767	0.39506	0.20981	0.18682	0.17325	0.09519	0.02744
Equal Eigenvalues									
λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}
56284	56284	56284	56284	56284	56284	56284	56284	56284	56284

5.4.2 Pseudorandom

The pseudorandom dataset was utilized because it provided a venue to observe and measure the behavior of the estimator set when perfect knowledge about the relationship between the response variable and the explanatory variables was known. Benefiting from a laboratory-like setting, the exact population was specified so that there was no lingering uncertainty about the impact of omissions, measurement error, or sampling bias. If the true population parameters, the β vector, were known then how would each estimator perform relative to the others? That is, if an ideal estimator could be designed that could achieve the optimum mean squared error reduction, what would be the results? By choosing the X-matrix of observed predictor values, the β vector of coefficients, and the ε vector of residuals, precise knowledge about the relationship among them was known.

Of course, such a data set is devoid of economic content or of any other subsequent conjecture about the implications of the relationship or application of the results. Also, the true relationship between any single x_{ij} and the response variable y_i, as embodied in β_j, is not necessarily expressible mathematically. A stronger statement supported empirically is that the true relationship is not possible to express mathematically--what is actually observed is that often a single x_{ij} will occur with multiple values of y_i. For instance, families with essentially the same income (x_{ij}) will have extremely diverse expenditure patterns (y_i). Hence, the population model

$$y = X\beta ,$$

also known as the deterministic model, while appealing, is not useful. A functional approach to the problem is to defer to reality and specify the regression model

$$y = X\beta + \varepsilon$$

instead. Since the model can be completely determined by knowing three of the four unknowns and, mathematically, since it does not matter which three, X, β, and ε were selected leaving y to be calculated. The experiment required the initial conditions to be predetermined. The following conditions were set:

- 1000 observations for the population (in deference to our base 10 metric system).
- 3 regressors so that the outcomes could be graphed[10]
- a seed of 3, 7, or 11.
- sample sizes of 10 (again, in deference to our base 10 metric system).
- X and β were selected from uniform distributions and ε was selected from a normal distribution[11]. In this way the values derived were not so small nor so large as to be unwieldy.
- the experiment was repeated 5000 times.

The vector of responses y was then calculated as

---

[10]This restriction precluded the use of Stein-type rules because of the $k > 4$ rule, but there was much explanatory value in a graphical treatment.

[11]The SAS random number generator NORMAL had an upward bias so it was necessary to fully scale the residual vector in some experiments. This is explained further in Section 6.2.

$$y = X\beta + \varepsilon .$$

The sample size of 10 was chosen because it is the small sample properties of estimators that is uncertain while asymptotic properties are better understood. Since economic experiments are often expensive to conduct and since much of the data gathered is historical, an estimation approach that performs well in small samples is of great value.

The data was unscaled; that is, eigenvectors and eigenvalues were derived from the product or second moment matrix $X'X$. However, since the values were generated from the same seed and the same distribution, the $X'X$ matrix bears a resemblance to a fully scaled correlation matrix (centered to a mean of 0 and variances in standard units) which, according to Jackson (1990), is widespread in some disciplines. He further points out that some computer packages only permit decomposition using the correlation matrix. In any case, the data points generated in the pseudorandom data set are unencumbered by labels so that attention can be focused on the comparative statistics.

Successive draws from the population was accomplished by generating 10 random integers from 1 to 1000 and selecting those observation numbers from X. This X matrix, unlike the Gorman and Toman X matrix, was not fixed in repeated samples but was permitted to vary from draw to draw. The nature of the dependency between X and ε covered in section 5.3 was dealt with by varying the value of the seed used to generate the random vectors. To simulate dependence, the same seed value was

used for X, β, and ε. To insure independence, a different seed was used for each, the prime numbers 3, 7, and 11 respectively.

As with the Gorman and Toman data, the MSE was calculated as

$$MSE^q = \left(\beta^q - \beta\right)' \left(\beta^q - \beta\right)$$

and subsequently averaged over the 5000 experiments.

Little multicollinearity was present in this data; the condition number of the product matrix was 3.29 and for the correlation matrix was .12.

TABLE 5.3
EIGENVALUES FOR THE PSEUDORANDOM DATASET

Product Matrix (X'X) Eigenvalues		
λ_1	λ_2	λ_3
835.722	95.097	77.3592
Correlation Matrix Eigenvalues		
λ_1	λ_2	λ_3
1.117	.979	.903

5.5 Empirical PER Shrinkage Estimators

As noted earlier, since θ_j and σ^2 are unknown an optimal estimator is problematic; nevertheless, it is possible that the orthogonality property of the principal components coefficients, θ_j's, might enable the OLS estimates θ_j's to provide robust substitutes that may help retain "near" optimality for

the ADHOC, INDIV and MATRIX estimators. By Slutsky's Theorem[12], any estimator that replaces θ with θ^o where θ_j^o is the ordinary least squares estimator of the principal components coefficients and the j^{th} element of the $k \times 1$ vector

$$\theta_j^o = (A'X'XA)^{-1} A'X'y$$

and where A is the matrix whose columns are eigenvectors of $X'X$ and $\hat{\sigma}^2$ is any consistent estimator of σ^2 such as

$$\hat{\sigma}^2 = \left(y - X\beta^o\right)' \left(y - X\beta^o\right) \Big/ (n - k)$$

is a consistent estimator of a_{ij}.

Empirical ADHOC

$$a_{ij}^a = \frac{\theta_j^{o2}}{\theta_j^{o2} + \frac{\hat{\sigma}^2}{\lambda_j}} a_{ij}$$

Empirical INDIV

$$a_{ij}^e = \frac{\lambda_j \theta_j^o \sum_{j=1}^{K} a_{ij} \theta_j^o}{\left[\sum_{j=1}^{K} \lambda_j \theta_j^{o2} + \hat{\sigma}^2\right]}$$

---

[12]Slutsky Theorem: if $\text{plim}\theta^* = \theta$ and $g(\theta^*)$ is a continuous function of θ^* then $\text{plim}g(\theta^*) = g(\theta)$. For a proof see S.S. Wilks, 1962. Mathematical Statistics. (Wiley, New York), pp. 102-103.

Empirical MATRIX $A^m = A\theta_j^o\theta_j^{o\prime}\left[\theta_j^o\theta_j^{o\prime} + \hat{\sigma}^2\ \Lambda^{-1}\right]^{-1}$

CHAPTER 6

MONTE CARLO RESULTS

6.1 The Gorman and Toman Dataset

Because the Gorman and Toman data has ten regressors, trying to come to conclusions about the individual parameter estimates is unwieldy. Thus, most attention will be directed towards the MSE results in this section. First an analysis of MSE results using the product matrix will be discussed, then the results of the standard errors will be examined, and finally, a discussion of performances using equal eigenvalues (no multicollinearity) will take place.

For convenience the following abbreviations are used:

OLS	ordinary least squares	PCM	principal components Mundlak
PEM	principal elements Mundlak	A	ADHOC
I/M	INDIV/MATRIX	PPO	positive part OLS
PPCM	positive part PCM	PPEM	positive part PEM
PPA	positive part ADHOC	PPI/M	positive part INDIV/MATRIX
DBO	Dey and Berger OLS	DBCM	Dey and Berger PCM
DBEM	Dey and Berger PEM	DBA	Dey and Berger ADHOC
DBI/M	Dey and Berger INDIV/MATRIX		

6.1.1 Product Matrix Eigenvalues

Table 6.1 contains the summary statistics for MSE calculations of each of the 18 estimators using the usual eigenvalues obtained from the product matrix $X'X$

Product Matrix (X'X) Eigenvalues

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}
528363	32899	951	362	162	98	5.80	1.33	.16	.007

TABLE 6.1
AGGREGATE STATISTICS FOR THE GORMAN AND TOMAN DATASET

Method	MSE	Standard Error	Minimum	Maximum
DBA	1304	396	123	6345
PPA	1377	52	437	3352
PPI/M	1469	819	103	9637
DBCM	1475	897	1358	14431
PPEM	1625	1103	430	12616
PCM	1865	1683	1341	14431
PEM	1936	1772	1293	14472
I/M	2154	1500	75	10871
PPO	2205	1661	74	12996
OLS	3229	2182	105	14438
A	3273	2071	134	12231
DBEM	3412	3116	92	25467
DBI/M	3562	2072	79	13025
DBO	5580	2813	104	18479
PPCM	7.9E24	1.03E26	1341	3.5E27

Since INDIV and MATRIX produced the same results they are combined so there are 15 rows. For convenience, the rankings are by MSE. MSE statistics by themselves are contained in Table 6.2 which is organized so that the columns represent the base estimation procedures. Three characteristics of the average value standout--(1) variations of PP clearly dominate the untreated forms, (2) dB variations do not dominate the untreated versions, and (3) the new PER forms ADHOC and INDIV/MATRIX produced small MSE's.

TABLE 6.2
MEAN SQUARED ERROR: GORMAN AND TOMAN

ADHOC	I/M	PE	PC	OLS
A	DBI/M	DBEM	PPCM	DBO
3273	3562	3412	7.9E24	5580
PPA	I/M	PEM	PCM	OLS
1377	2154	1936	1865	3229
DBA	PPI/M	PPEM	DBCM	PPO
1304	1469	1625	1475	2205

The standard errors are summarized in Table 6.3, again with the columns corresponding to the type of base estimator. What is particularly impressive is that the PP and DB forms of the new PER shrinkage estimator ADHOC have much tighter distributions; only the untreated ADHOC estimator is not situated with the others. As well, the PP form of ADHOC has a very tight error distribution. The new PER forms ADHOC and INDIV/MATRIX, with 4 of the 6 smallest values, are able to concentrate more of the distribution close to the desired mean than are alternative forms.

Otherwise, there is no clear domination tendency between DB and PP, although both outperform the untreated forms, having 6 of the 7 smallest values.

TABLE 6.3
STANDARD ERROR OF MSE: GORMAN AND TOMAN

ADHOC	I/M	PE	PC	OLS
A	DBI/M	DBEM	PPCM	DBO
2071	2072	3116	1.03E25	2813
DBA	I/M	PEM	PCM	OLS
396	1500	1772	1683	2182
PPA	PPI/M	PPEM	DBCM	PPO
52	819	1103	897	1661

Table 6.4 summarizes the maximum value of each estimator. Notice that the 5 lowest values are all forms of ADHOC or INDIV/MATRIX, even the untreated versions; the two new PER shrinkage methods have implications of being minimax. Also, the positive-part variants tend to accumulate at the lowest maximums.

TABLE 6.4
MAXIMUM MSE: GORMAN AND TOMAN

ADHOC	I/M	PE	PC	OLS
A	DBI/M	DBEM	PPCM	DBO
12231	13025	25467	3.5E27	18479
DBA	I/M	PEM	PCM	OLS
6345	10871	14472	14431	14438
PPA	PPI/M	PPEM	DBCM	PPO
3352	9637	12616	14431	12996

Notice in Table 6.5 containing minimum values that 3 of the 5 lowest minimum values are all forms of INDIV/MATRIX. A striking statistic is the high minimums for all three forms of PCR, accompanied by the untreated PER estimator. Paradoxically, a minimum that approaches the target of 0 also inflates the variance so this statistic is not as interesting as the others.

TABLE 6.5
MINIMUM MSE: GORMAN AND TOMAN

ADHOC	I/M	PE	PC	OLS
PPA	PPI/M	PEM	DBCM	OLS
437	103	1293	1358	105
A	DBI/M	PPEM	PPCM	DBO
134	79	430	1341	104
DBA	I/M	DBEM	PCM	PPO
123	75	92	1341	74

In summary, the new PER forms ADHOC and INDIV/MATRIX perform very well in comparison to the other estimators when evaluated by mean

squared error and standard error. Further, examination of the maximum MSE's intimates that ADHOC and INDIV/MATRIX may be minimax. Simulations on the Gorman and Toman dataset are very encouraging for the new PER forms.

6.1.2 The Beta Vector

Turning attention to the estimates of the beta vector, the problems caused by the dataset's dimension and the large number of estimation techniques precludes the same graphical approach as will be used for the pseudorandom dataset. Instead, three methods will be used. One focuses on the actual parameter estimate and a second is concerned with the standard error of each component of the beta vector. Third, a useful but rather simple way to summarize the results within a statistical category is -- by each β_i -- to assign a score of 1 if a method had the most desirable value, a 2 if it had the next most desirable, and so on, and then compute the average rank order value. Low values imply that an estimator performed well overall.

Table 6.6 lists the average estimates of β_i in separate categories where the cell containing the true beta is in bold for easier location. As expected, since OLS is an unbiased estimator in almost every instance OLS, PPO, and DBO are relatively close to BETA, with some instability noticeable for the last 3 or 4 β_i. Among the remaining estimators, INDIV/MATRIX variations are relatively close to BETA except that the positive part variant is

consistently farther away than either of the other two. Table 6.7 lists the standard errors of β_i in separate categories.

TABLE 6.6

AVERAGE BETA SUMMARY: GORMAN AND TOMAN

	PPC	PPI/M	PPA	DBC	PC	A	DBA	PPO	I/M	DBI/M				PPE	DPE	PE
β1	-8.90E+19	-6.7	-5.7	-0.6	-0.6	-0.1	-0.1	2.1	2.4	3.6				6.9	12.6	13.5
	DPE	PE	PPE				DBI/M	PPO	I/M	DPC	PC	DBA	A	PPC	PPI/M	PPA
β2	-8.3	-8.1	-7.2				-4.5	-3.9	-3.9	-1.8	-1.8	-1.7	-1.7	-1.6	-1.3	-1
	PPO				DPE	PPO	I/M	PE	PPI/M	PPE	DBA	A	DPC	PC	PPA	PPC
β3	-106				-257.7	-209	-201.6	-106	-69.5	-43.8	-36.1	-34.6	-2.6	-2.6	-1.4	7.94E+21
	PPC	DPE	PE	PPE	DBO				PPO	I/M	DBA	A	DPC	PC	PPI/M	PPA
β4	-7.60E+20	-49	-32.1	-25.9	-16.2				-11	-10.9	-3	-2.9	-0.6	-0.6	-0.5	1.7
	PPC				DBI/M	PPO	I/M	PPI/M	PE	PPE	DBA	A	DPC	PC	PPA	DPE
β5	-1.06E+21				-1169.1	-863	-830	-292	-186.5	-159	-145	-140	-14.3	-14.3	-6	648.6
	PPC	PPA	DPC	PC	DPE	A	DBA	PPI/M	PPE	PE	I/M	PPO	DBI/M			
β6	-9.63E+20	3.2	6.3	6.3	41.9	48.4	49.1	80.8	80.8	124.4	225.8	232.3	276.8			
	DPC	PC				CLS	PPI/M	A	DBA	PPA	I/M	PE	DBO	DBI/M	PPE	DPE
β7	-1.7	-1.7				-0.9	-0.9	-0.9	-0.8	-0.7	-0.6	-0.3	-0.2	-0.1	-0.1	0.7
	PPC	DPC	PC	PPA	A	DBA	PPI/M	I/M	PPO	DBI/M	PPE	PE				DBO
β8	-2.38E+18	0.9	0.9	3.6	4.3	4.4	9.5	16.3	16.6	18.5	18.6	20.8				28.2
	PPE	PE	DPC	PC	PPC				CLS	DBA	PPA	PPO	DBO	I/M	DBI/M	PPI/M
β9	-0.5	-0.5	-0.5	-0.5	-0.4				-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.1	-0.1
	DPE	PE	PPE	A	DBA	DPC	PC	PPC	I/M				PPO	DBI/M	DBO	PPI/M
β10	-4.5	-3.2	-2.2	-0.4	-0.4	-0.4	-0.4	-0.3	0.1				0.6	0.8	1.2	1.9

TABLE 6.7
STANDARD ERROR BETA SUMMARY: GORMAN AND TOMAN

	PPA	DBA	A	PC	DBC	PPI /M	I /M	PPO	DBI/M	OLS	PE	DBE	PPE	DBO	PPC
β1	7.48	12.20	12.22	13.08	13.08	20.22	30.67	32.36	37.25	46.32	47.37	50.35	51.65	56.62	4E+21
	PPA	DBA	A	PPC	PC	DBC	PPI/M	I/M	PPO	DBI/M	OLS	PE	PPE	DBE	DBO
β2	2.85	3.02	3.02	3.04	3.04	3.04	7.11	8.47	9.26	11.03	12.76	13.15	13.22	13.68	16.76
	PPA	PC	DBC	A	DBA	PPE	PPI/M	PE	I/M	PPO	OLS	DBI/M	DBE	DBO	PPC
β3	25.55	61.04	61.04	109.31	111.53	249.38	263.14	311.62	520.86	548.75	781.55	840.68	892.98	1269.41	9.60E+22
	PC	DBC	PPA	A	DBA	PE	PPE	PPI/M	I/M	PPO	DBE	OLS	DBI/M	DBO	PPC
β4	7.62	7.62	17.35	18.83	19.30	58.01	62.29	67.19	88.38	93.20	117.97	133.36	142.83	216.05	1.02E+22
	PPA	PC	DBC	A	DBA	PPI/M	PPE	PE	I/M	PPO	OLS	DBI/M	DBE	DBO	PPC
β5	129.63	427.52	427.52	577.07	587.93	1262.67	1304.63	1640.71	2513.22	2652.87	3799.81	4015.21	4209.10	6075.67	2.47E+22
	PPA	PC	DBC	A	DBA	PPI/M	I/M	PPE	PPO	DBI/M	PE	OLS	DBO	DBE	PPC
β6	24.81	76.46	76.46	78.37	78.64	149.38	1233.56	238.23	257.88	328.80	331.39	347.14	491.67	531.73	5.56E+22
	PPA	PPC	PC	DBC	A	DBA	PPI/M	I/M	PPO	DBI/M	PPE	OLS	PE	DBE	DBO
β7	2.79	3.12	3.19	3.19	3.41	3.43	7.92	11.62	12.32	17.00	17.51	17.59	17.68	22.32	25.88
	PPA	A	DBA	PC	DBC	PPI/M	I/M	PPO	PE	PPE	OLS	DBI/M	DBE	DBO	PPC
β8	6.35	6.76	6.80	7.82	7.82	17.42	21.28	22.78	28.29	29.40	31.76	31.89	35.13	47.91	1.68E+20
	PC	DBC	PPC	PPA	A	DBA	PPI/M	PE	PPE	I/M	PPO	OLS	DBI/M	DBE	DBO
β9	0.19	0.19	0.21	0.26	0.30	0.30	1.05	1.13	1.16	1.28	1.46	1.93	2.05	2.30	3.11
	PPC	PC	DBC	PPA	A	DBA	PPI/M	I/M	PPO	PE	PPE	DBE	OLS	DBI/M	DBO
β10	2.92	2.98	2.98	3.04	3.29	3.34	9.79	12.09	13.09	16.51	16.97	17.41	18.28	18.64	28.22

Another way to simplify is to average the standard errors of each estimator over all betas: see Table 6.8a. ADHOC estimators perform very well because they have smallest errors across all parameters. Considering the treatments, there is no clear pattern. Two versions of PCM produced impressive results (but be reminded that pre-test estimators have unbounded risk so their performance is more sensitive to the sample information than the others).

TABLE 6.8a
BETA SUMMARY GORMAN AND TOMAN:
AVERAGE STANDARD ERROR OVER ALL BETAS

PPA	A	DBA	PCM	DBCM	PPI/M	PPEM	PEM
22.01	81.26	82.65	158.86	158.86	180.59	234.69	305.51
I/M	PPO	OLS	DBI/M	DBEM	DBO	PPCM	
344.14	364.40	519.05	544.54	587.52	823.13	1.7E24	

Table 6.8b is the average rank order of standard errors based on Table 6.7. For instance, PP ADHOC receives 1's for β_1, β_2, β_3, β_5, β_6, β_7, β_8, a 3 for β_4, and 4's for β_9, β_{10}. Thus, its average rank is 1.8. Each remaining beta is ordered and averaged in the same fashion. When the estimators are characterized this way, ADHOC has 3 of the 5 smallest values, ADHOC and INDIV/MATRIX together have 5 of the 7 smallest values, and even two PCM forms also do well (remember, pre-test estimators have unbounded risk). Among the types of treatments, the three largest values are DB forms so PP forms have an edge.

TABLE 6.8b
BETA SUMMARY GORMAN AND TOMAN:
AVERAGE STANDARD ERROR RANK OVER ALL BETAS

PPA	PCM	DBCM	A	DBA	PPI/M	I/M	PPO
1.8	2.6	3.6	3.9	4.5	6.7	8.2	9.3
PPEM	PEM	PPCM	OLS	DBI/M	DBEM	DBO	
9.5	9.6	10.0	11.4	11.5	13.0	14.3	

Overall, if the performance criterion is solely the accuracy of an estimator as manifested by just the mean estimate, unbiased OLS versions would be preferred, INDIV/MATRIX is a close second, and the relative performance of the remaining ones is unclear. If the performance standard instead is only the distribution of estimates, ADHOC would be the estimator of choice while INDIV/MATRIX follows closely. Combining the two standards, accuracy and distribution, INDIV/MATRIX does well and could be characterized as the wisest choice.

6.1.3 Clustering of the Beta Estimates

This section presents a non-graphical, simplified way to examine the clustering of parameter estimates around the true parameter. Even if biased estimators have a reduced mean squared error over OLS, if the distribution is located so far from the target parameter that it has little chance to provide a precise estimate, the estimator may be undesirable. In other words, it may

be desired that the envelope defined by the scatter of point estimates have the opportunity to contain an exact or near-exact estimate.

To define 'precise' for this analysis, a range around the true beta was arbitrarily chosen as 100. For instance, the β_i estimates were sorted by magnitude, the true beta located, and then the 50 values on either side of it were selected. Table 6.9 contains that summary.

The three forms of ADHOC concentrate a large number of values close to the true parameter in 5 of the betas, β_1, β_4, β_7, β_9, and β_{10}. In fact, for β_7, β_9, and β_{10} it has the three largest values and in β_1 and β_4 it has 2 of the three largest. In β_2 there are a many close values, too.

Unlike ADHOC forms, OLS forms never have the largest number of values. Only in β_3, β_5, and β_6 did it produce a large number of precise estimates while in β_2, β_4, and β_8 there were many close values.

INDIV/MATRIX forms are very consistent across all betas, doing particularly well in β_2, β_5, β_6, and β_8 with the largest values in addition to other large values. Also, they have many large values in β_1, β_3, β_7, β_9, and β_{10}, too.

Thus, not only did the PER forms ADHOC and INDIV/MATRIX estimators tend to produce small MSE's and small standard errors, they also located precise point estimates near the true parameter as often or more often than did OLS and its PP and DB variants.

TABLE 6.9

CLUSTERING OF BETAS GORMAN AND TOMAN

	A	DBA	PPI/M	DBE	PPO	PEM	I/M	DBI/M	PPA	PPEM	DBO	DBCM	PPCM	OLS	PCM
β1	16	13	10	9	8	5	5	5	3	3	3	0	0	0	0
	I/M	PPO	DBEM	DBI/M	DBA	A	OLS	PPA	PPEM	PPI/M	PEM	DBO	PPCM	DBCM	PCM
β2	14	10	8	8	7	7	5	4	3	3	3	3	1	0	0
	DBE	PPO	I/M	OLS	PEM	DBI/M	DBO	PCM	DBC	PPI/M	DBA	PPEM	A	PPA	PPCM
β3	16	15	11	11	7	4	4	3	3	3	2	2	1	0	0
	DBA	PPEM	A	PEM	DBEM	OLS	PPO	I/M	PPA	PPI/M	DBCM	DBO	DBI/M	PPCM	PPC
β4	20	17	14	11	9	7	6	4	4	2	0	0	0	0	0
	I/M	PPO	OLS	DBEM	DBI/M	PPI/M	DBO	DBA	A	DBCM	PPCM	PPEM	PPA	PEM	PCM
β5	17	16	10	9	7	6	3	2	1	0	0	0	0	0	0
	I/M	OLS	DBI/M	DBO	PPO	PEM	PPI/M	DBEM	PPEM	A	DBA	DBCM	PPCM	PPA	PCM
β6	12	11	9	9	8	7	6	4	4	2	1	0	0	0	0
	PPA	A	DBA	PPCM	PPI/M	PCM	DBCM	DBO	PPO	I/M	DBE	DBI/M	OLS	PPEM	PEM
β7	21	14	12	10	5	4	4	4	4	3	2	2	2	2	1
	DBI/M	PPO	I/M	DBE	PEM	PPEM	PPI/M	OLS	DBO	DBA	DBCM	PPA	PPCM	A	PCM
β8	10	10	10	8	7	7	6	6	5	1	1	1	1	1	1
	PPA	DBA	A	PPCM	PPEM	I/M	DBCM	PPI/M	PCM	PPO	PEM	OLS	DBI/M	DBO	DBEM
β9	31	14	11	10	5	4	3	3	3	2	2	2	1	1	0
	A	PPA	DBA	PPI/M	DBEM	PPEM	PPO	OLS	DBI/M	I/M	PEM	DBO	DBCM	PPCM	PCM
β10	22	20	15	7	5	5	4	4	3	3	1	0	0	0	0

6.1.3 Equal Eigenvalues

Table 6.10 contains the summary statistics for MSE calculations of each of the 18 estimators when the eigenvalues are all equal.

Equal Eigenvalues

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}
56284	56284	56284	56284	56284	56284	56284	56284	56284	56284

Since INDIV and MATRIX produced the same results they are combined as done previously so there are 15 rows. For convenience, the rankings are by MSE.

TABLE 6.10
AGGREGATE STATISTICS FOR THE
GORMAN AND TOMAN DATASET: EQUAL EIGENVALUES

Method	MSE	Standard Error	Minimum	Maximum
A	17.64	2.38	10.31	128.99
I/M	3229.36	2182.34	105.90	14438.39
PCM	3229.37	2182.34	105.90	14438.39
OLS	3229.37	2182.34	105.90	14438.39
PEM	3229.37	2182.34	105.90	14438.39
DBA	3229.39	2182.32	104.48	14438.38
DBI/M	3229.57	2182.21	105.97	14438.39
DBEM	3229.58	2182.21	105.95	14438.39
DBO	3229.58	2182.21	105.95	14438.39
DBCM	3229.59	2182.21	106.09	14438.40
PPO	3229.58	2182.21	105.95	14438.39
PPCM	3229.58	2182.21	105.95	14438.39
PPEM	3229.58	2182.21	105.95	14438.39
PPA	3229.58	2182.21	105.95	14438.39
PPI/M	3229.58	2182.21	105.95	14438.39

It is expected that there would be little difference among the estimators in this scenario because the absence of dependencies means there is no way for an estimator to determine which elements to impact. Essentially, that is what occurred for the untreated and DB versions since the minor differences are likely due to rounding errors. Thus, the performance of ADHOC in table 6.10 is particularly impressive. It dominates all other estimators resoundingly in all statistics!

6.2 The Pseudorandom Dataset

Analytic properties of estimators are often derived under convenient assumptions such as normally distributed residuals with a mean of zero and a variance 1 or independence between the residuals and the explanatory variables. This portion of the dissertation examines the estimator set under those conditions as well as unkind (and likely more realistic) circumstances where those assumptions are explicitly violated. Five sets of simulations were performed with this dataset.

In Section 6.2.1 and 6.2.2 analysis is on a dataset that has correlated explanatory variables and residuals in addition to biased residuals (although they are normally distributed) under a low multicollinearity structure. Next, Section 6.2.3 explores the impact of the independence/dependence assumption discussed in Section 5.3 using the eigenvalues of the product matrix (the low multicollinearity scenario). Finally, Section 6.2.4 examines the independence/dependence assumption when multicollinearity is high.

In all simulations, 5000 random samples of size 10 were taken from a population of 1000. Summary statistics on the MSE and the beta vector were collected for the each of the estimators. The MSE of each was compared--the mean, the standard error, the minimum, and the maximum. Additionally, how closely each approximated the beta vector was analyzed--the mean, the standard error, the minimum, and the maximum.

6.2.1 Mean Squared Error: UNKIND Data

Aggregate MSE statistics for each of the estimators is reported in Table 6.11 and subsets of that table follow.

TABLE 6.11
AGGREGATE STATISTICS FOR THE UNKIND
PSEUDORANDOM DATASET

Estimator	MSE	Standard Error	Minimum	Maximum
I/M	7.56	5.39	0.47	58.94
ADHOC	12.69	6.00	1.10	60.95
PEM	13.62	8.28	0.43	84.64
PCM	13.73	8.16	0.52	84.64
OLS	13.96	8.05	0.53	84.64

The first notable statistic is that the mean MSE of OLS exceeds every other estimator. Table 6.12 ranks those results and identifies INDIV/MATRIX as having the lowest MSE on average, ADHOC has the next lowest, PEM nudges PCM, and OLS has the highest value.

TABLE 6.12
MEAN SQUARED ERROR UNKIND
PSEUDORANDOM DATASET

I/M	ADHOC	PEM	PCM	OLS
7.56	12.69	13.62	13.76	13.96

Table 6.13 contains rankings by standard errors. Consistent with the rankings by maximum value, the order is INDIV/MATRIX, ADHOC, OLS, PCM, and PEM.

TABLE 6.13
STANDARD ERROR OF MSE UNKIND
PSEUDORANDOM DATASET

I/M	ADHOC	OLS	PCM	PEM
5.39	6.00	8.05	8.16	8.28

It is also interesting to examine the minimum (Table 6.14) and maximum (Table 6.15) MSE values. Since the minimums vary little compared to the maximums, judgment by this statistic is not as noteworthy but the ranking is: INDIV/MATRIX achieved the lowest, PEM attains a minimum almost as low as INDIV/MATRIX, while OLS, PCM and OLS have nearly equal values. Comparing maximums, INDIV/MATRIX has the lowest, ADHOC follows, and the remaining three methods tie for the highest maximum.

TABLE 6.14
MINIMUM MSE UNKIND PSEUDORANDOM DATASET

PEM	I/M	PCM	OLS	ADHOC
0.43	0.47	0.52	0.53	1.10

TABLE 6.15
MAXIMUM MSE UNKIND PSEUDORANDOM DATASET

I/M	ADHOC	PEM	PCM	OLS
58.94	60.95	84.64	84.64	84.64

Clearly, INDIV and MATRIX are more desirable based upon any single of these statistics and in the aggregate as well. ADHOC dominates the remaining estimators except for the minimum value, but there is little difference among all six on this basis anyway. The maximum values of ADHOC and INDIV/MATRIX are similar and suggest that they may be minimax in this setting. Labeling any one of the remaining three as better than the other two is arguable, but it is obvious that they are dominated by the three new PER shrinkage estimators.

A graphical summary of the simulations is presented next in Figure 6.1. Assume that the MSE statistic is distributed as a χ^2 with $v = (n - k - m)$ degrees of freedom where n is the sample size, k is the number of restrictions imposed by the column dimension of X, and m is the number of linear restrictions imposed by the estimation method. In these simulations $n = 10$, $k = 3$, and $\max(m = 2)$ (since if $m = 3$ all elements of A would be 0). As m increases causing v to decrease, the distribution compresses so that for a given MSE more of the probability density is located close to the optimum value of 0.

Theoretically, determining the number of linear restrictions each estimator places on the eigenvector matrix A can be found merely by counting the number of 0 eigenvalues, but practically this necessitates

deciding what value is so small that it is essentially equal to 0. That can be avoided by assuming the worse case ($m = 2$) and then making comparisons.

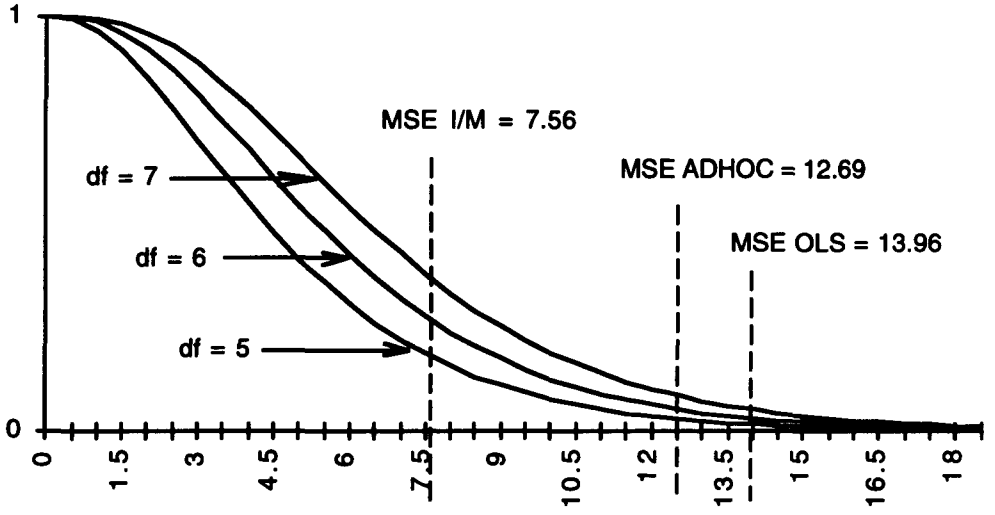


FIGURE 6.1
CHI-SQUARE DISTRIBUTIONS
OF MEAN SQUARED ERROR: UNKIND PSEUDORANDOM DATASET
population size = 1,000 sample size = 10 experiment size = 5,000

However, since MSE is also changing, a reasonable way to compare the estimator set is to find out how much of the distribution is contained between the mean estimate and 0. Table 6.16a reports the probability densities close to zero under different assumptions about the degrees of freedom. Since OLS imposes no additional restrictions than k, the MSE of OLS possesses a χ^2 distribution with 7 degrees of freedom, whereas the MSE of INDIV/MATRIX has as few as 5 degrees of freedom. Regardless of

the distribution, a lower probability is a positive characteristic because the MSE value is closer to 0. For this experiment, for OLS 94.8% of the probability density is left of the mean value 13.96. Assuming that ADHOC and INDIV/MATRIX have only 5 degrees of freedom, for ADHOC 97.4% and for INDIV/MATRIX only 81.8% is to the left of their mean 12.69 and 7.56, respectively. If the most severe assumption is relaxed, with 6 degrees of freedom ADHOC improves to 95.2% and INDIV/MATRIX to 72.8%.

TABLE 6.16a
P(0 < X < MSE)
UNDER DIFFERENT RESTRICTION ASSUMPTIONS

Method	I / M	ADHOC	PEM	PCM	OLS
MSE	7.56	12.69	13.62	13.73	13.96
		df = 7,	m = 0		
P(0<X<MSE)	.627	.920	.942	.944	.948
		df = 6,	m = 1		
P(0<X<MSE)	.728	.952	.966	.967	.970
		df = 5,	m = 2		
P(0<X<MSE)	.818	.974	.982	.983	.984

Another way to evaluate the outcome of the experiment is to calculate the distance of each mean MSE estimate from 0 in terms of its standard deviation s[13] as listed in Table 6.16b. By that standard, the OLS value is 3.73 s from 0 and, with 5/6 degrees of freedom, respectively ADHOC and INDIV/MATRIX are 4.01/3.67 s and 2.39/2.18 s from 0.

[13]The standard deviation of a $\chi^2 = \sqrt{Var(\chi^2)} = \sqrt{2v}$.

TABLE 6.16b
DISTANCE FROM 0 TO MEAN MSE
IN STANDARD DEVIATIONS s

Method	Standard	I/M	ADHOC	OLS
MSE	Error	7.56	12.69	13.96
df	s		# of s	
7	3.74	2.02 s	3.39 s	3.73 s
6	3.46	2.18 s	4.01 s	-
5	3.16	2.39 s	3.67 s	-

6.2.2 The Beta Vector

Turning to an evaluation of the beta vectors, refer first to Table 6.17 and the corresponding Figure 6.2.

TABLE 6.17
AGGREGATE STATISTICS FOR BETA 1

Estimator	Mean	Standard Error	Minimum	Maximum
ADHOC	2.191	3.622	-22.71	38.28
True Beta	2.608			
I/M	3.072	5.608	-23.90	44.07
PCM	5.495	8.982	-32.32	63.51
PEM	5.800	9.304	-32.32	63.51
OLS	5.936	9.552	-32.32	63.51

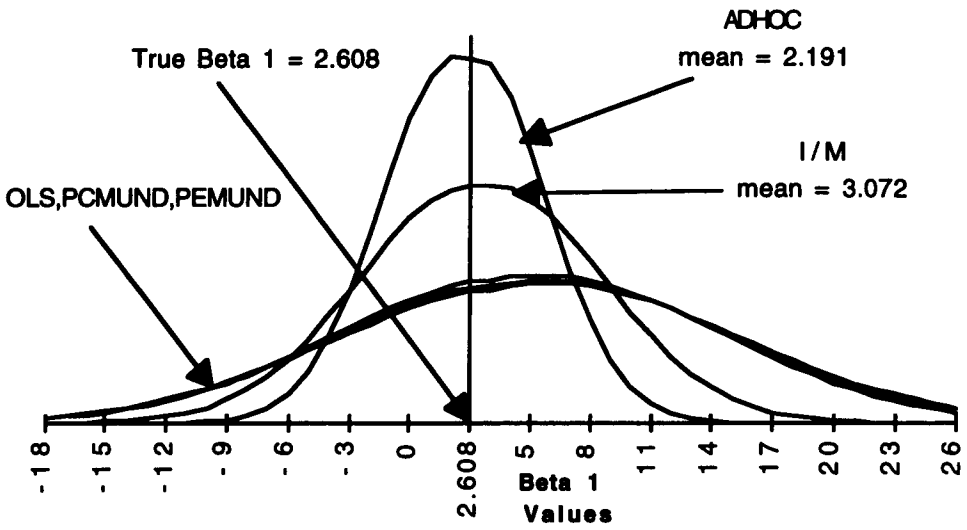


FIGURE 6.2
DISTRIBUTION OF BETA 1 FOR ALL ESTIMATORS

Noting that the true population beta was fixed as 2.608, once again ADHOC and INDIV/MATRIX are far superior to any of the other three in their ability to estimate β_1. Using the statistics in Tables 6.18 and 6.19, Figures 6.3a and 6.3b provide a visual comparison of the distributions of OLS with ADHOC and OLS with INDIV/MATRIX respectively. Not only is the average estimate of β_1 closer for both than for OLS, the graphs show the amount of each distribution within one OLS standard error of the true value 2.608. Whereas OLS has only 65.3% within that range, for ADHOC 99.1%, or nearly all of it, is within that range while INDIV/MATRIX accumulated 91% within that range.

TABLE 6.18
MEAN OF BETA 1

ADHOC	True Beta	I/M	PCM	PEM	OLS
2.191	2.608	3.072	5.495	5.800	5.936

TABLE 6.19
STANDARD ERROR BETA 1

ADHOC	I/M	PCM	PEM	OLS
3.622	5.608	8.982	9.304	9.552

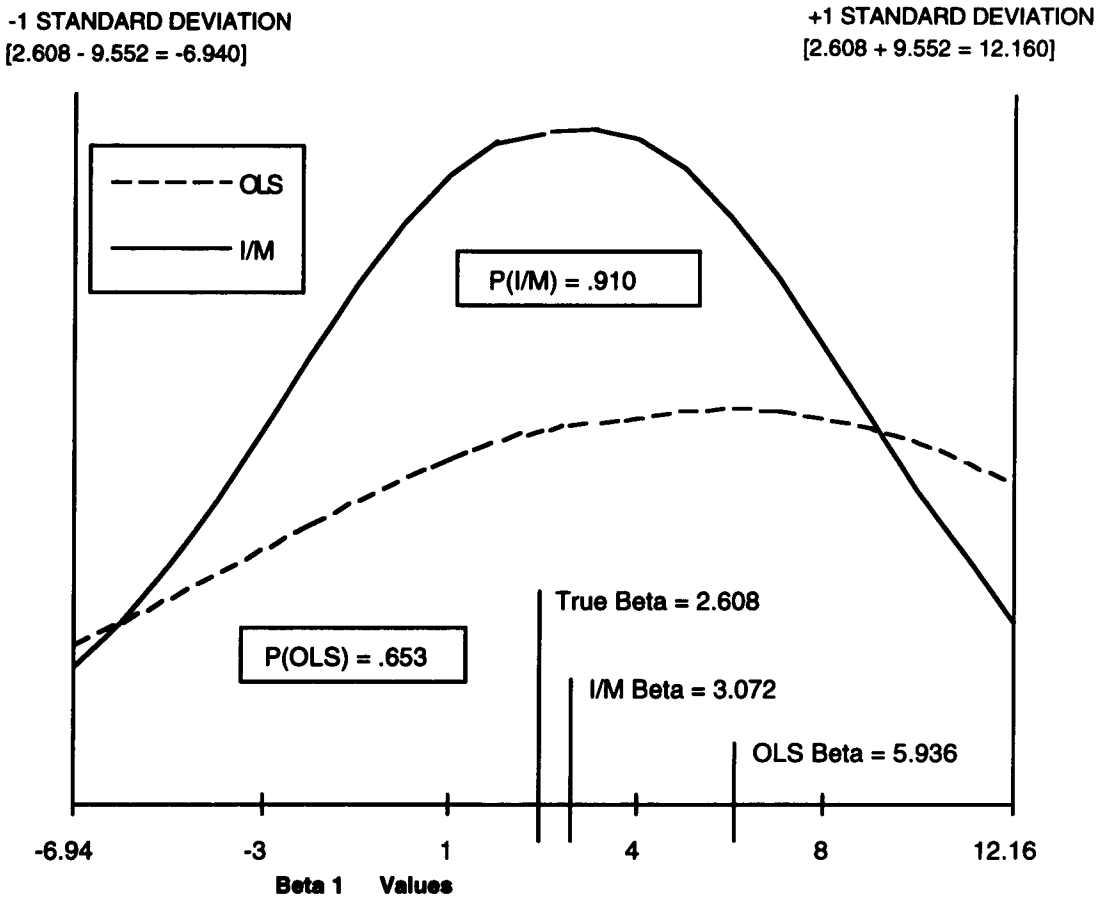


FIGURE 6.3a
BETA 1: PROBABILITY VOLUME WITHIN
1 OLS STANDARD ERROR OF THE TRUE MEAN

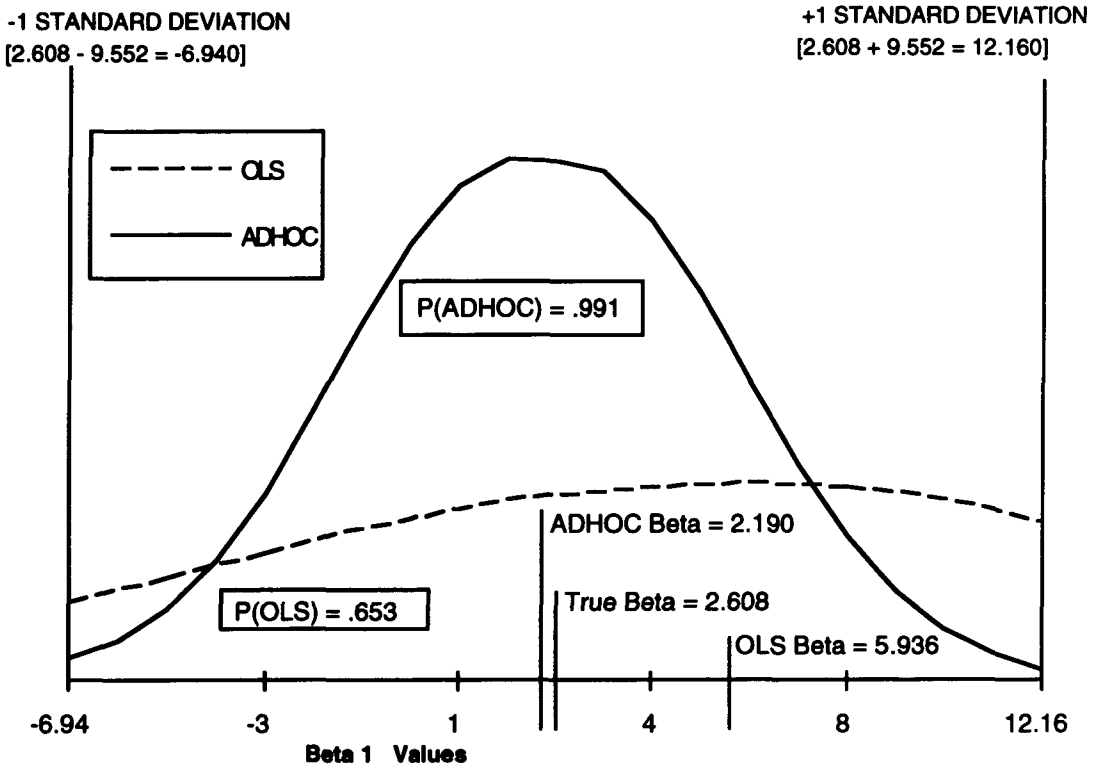


FIGURE 6.3b
BETA 1: PROBABILITY VOLUME WITHIN
1 OLS STANDARD ERROR OF THE TRUE MEAN

Tables 6.20, the minimum estimate of β_1, and 6.21, the maximum estimate, are worth noting. Both ADHOC and INDIV/MATRIX produce estimates of β_1 characterized by a smaller range that do their competitors.

TABLE 6.20 MINIMUM BETA 1
UNKIND PSEUDORANDOM DATA

ADHOC	I/M	PEM	PCM	OLS
-22.71	-23.90	-32.32	-32.32	-32.32

TABLE 6.21 MAXIMUM BETA 1
UNKIND PSEUDORANDOM DATA

ADHOC	I/M	PEM	PCM	OLS
38.28	44.07	63.505	63.505	63.505

While the results for the estimates of β_2 and β_3 obviously favor ADHOC and INDIV/MATRIX, the outcomes are not as striking as for β_1. Tables 6.22 and 6.23 contain the aggregate statistics that are represented in Figures 6.4 and 6.5, corresponding to β_2 and β_3 respectively.

An examination of the Table 6.22 and Figure 6.4 reveals that none were good point estimators of β_2, but ADHOC and INDIV/MATRIX had smaller standard errors once again.

Table 6.22 and Figure 6.5 represent β_3. Note that no estimator is in any sense "close" to β_3 but ADHOC and INDIV/MATRIX again exhibit the characteristic of small standard errors.

TABLE 6.22
AGGREGATE STATISTICS FOR BETA 2

Estimator	Mean	Standard Error	Minimum	Maximum
ADHOC	1.327	3.381	-30.841	24.521
I/M	1.490	5.608	-38.311	30.782
True Beta	2.199			
OLS	2.64	9.197	-62.159	43.591
PEM	2.673	8.963	-62.159	43.591
PCM	2.753	8.660	-62.159	43.591

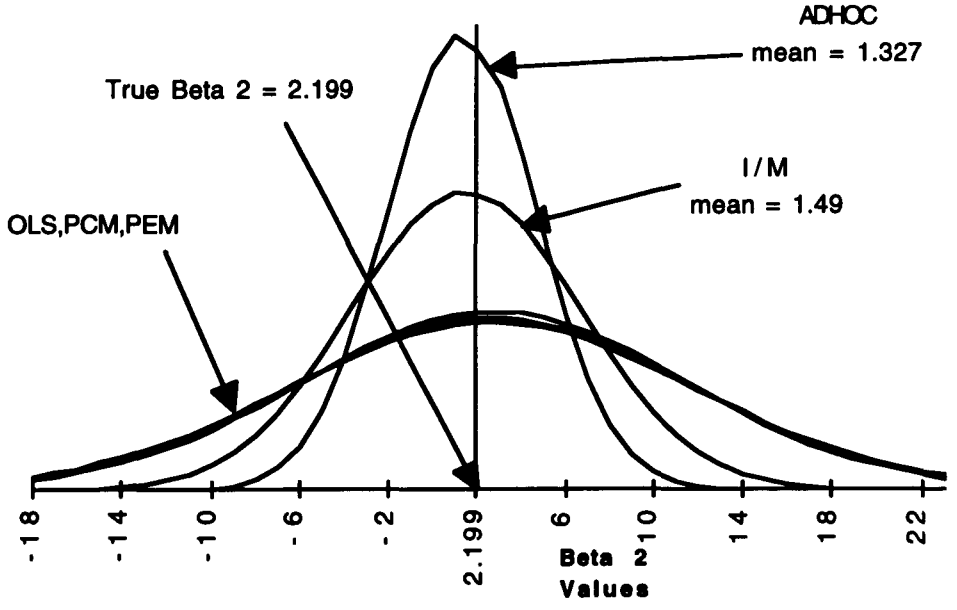


FIGURE 6.4
DISTRIBUTION OF BETA 2 FOR ALL ESTIMATORS

TABLE 6.23
AGGREGATE STATISTICS FOR BETA 3

Estimator	Mean	Standard Error	Minimum	Maximum
I/M	0.411	4.579	-41.847	29.386
ADHOC	0.712	2.931	-36.098	24.738
OLS	0.896	8.277	-60.207	37.494
PEM	1.028	8.053	-60.207	37.494
PCM	1.173	7.808	-60.207	37.494
True Beta	2.922			

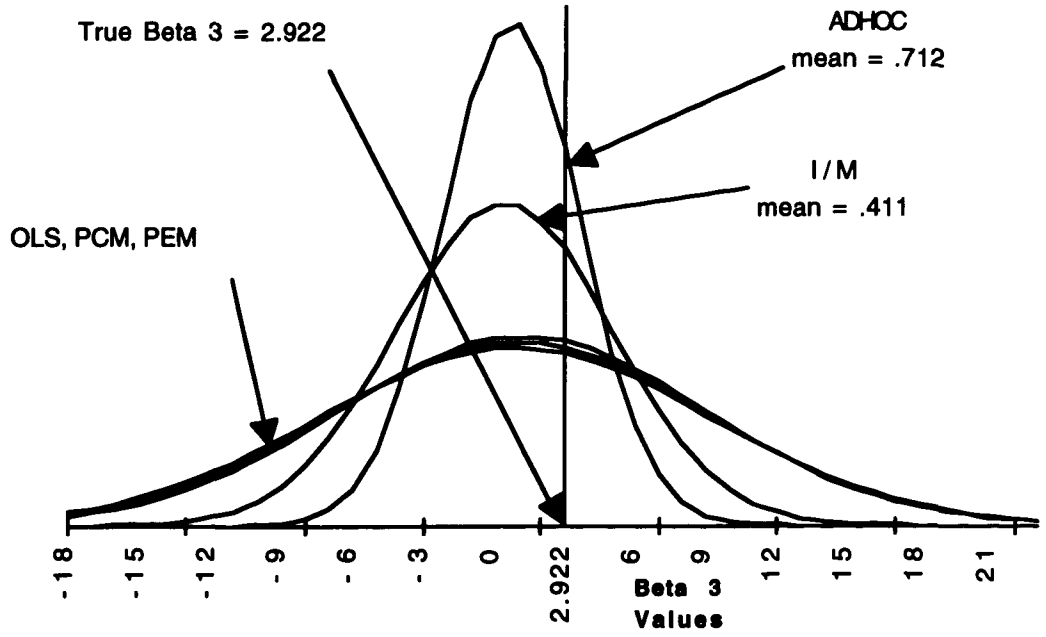


FIGURE 6.5
DISTRIBUTION OF BETA 3 FOR ALL ESTIMATORS

For all three parameters, ADHOC and INDIV/MATRIX have a tighter distribution about the population parameter. Indeed, it is this ability to

concentrate a larger amount of the probability distribution closer to the true parameter rather than the likelihood, however small, of estimating it precisely, that makes biased estimators more attractive than unbiased estimators.

6.2.3 Low Multicollinearity

Product Matrix (X'X) Eigenvalues

λ_1	λ_2	λ_3
835.722	95.097	77.3592

Separate analysis was done (LMI) so that independence between X and ε could be assumed (with different seeds for X, β, and ε) and then the independence assumption (LMD) was violated by using the same seed for each. Aggregate MSE statistics for each of the estimators are reported in Table 6.24.

TABLE 6.24
AGGREGATE STATISTICS
FOR THE PSEUDORANDOM DATASET LMI AND LMD

Independence				
Estimator	MSE	Standard Error	Minimum	Maximum
PEM	1.59	0.76	0.10	8.26
PCM	1.62	1.01	0.05	12.66
I/M	1.67	0.98	0.06	12.10
OLS	1.76	1.04	0.07	12.66
ADHOC	1.96	0.77	0.31	10.05
Dependence				
Estimator	MSE	Standard Error	Minimum	Maximum
PEM	1.13	0.61	0.05	9.57
PCM	1.46	0.87	0.09	10.03
I/M	1.55	0.87	0.08	9.85
OLS	1.59	0.91	0.12	10.06
ADHOC	1.76	0.67	0.39	7.03

By assumption, the discrepancy among the empirical mean squared errors of the competing estimators is not large and between assumptions, the relative position of the estimators is the same in both cases. The new PER form of the Mundlak estimator had the lowest MSE in both cases (easily the lowest for LMD) followed by PCM, INDIV/MATRIX, OLS (which were essentially even), and ADHOC was the largest with some separation from the others.

Turning attention to other statistics, the distribution of the estimates of both ADHOC and INDIV/MATRIX fare better than OLS while PEM again is impressive. The minimum estimates of MSE for four are very similar except

for the relatively large value of ADHOC. Offsetting ADHOC's large minimum is its very small maximum, followed by PEM, INDIV/MATRIX, PCM, and OLS.

The statistics just presented definitely favor the principal elements estimator PEM in terms of average MSE and standard errors combined--however, the unbounded risk characteristic of the pre-test estimator is a caution against accepting it too willingly.

The analysis of the beta vector here will not be as detailed as above with the UNKIND dataset. Referring to Table 6.25, under the independence assumption, INDIV/MATRIX obtained a very precise estimate of Beta while OLS, being unbiased, performed well relative to the others. When there is dependence between the residuals and the explanatory variables all methods obtained close estimates.

However, even in this low multicollinearity scenario OLS has larger variances than the others. While INDIV/MATRIX estimated the LMI Beta closely, the error was much larger than PCM, PEM, and ADHOC. For LMD where all estimators were quite close to the true parameter, ADHOC, PCM and PEM had significantly smaller errors.

TABLE 6.25
LOW MULTICOLLINEARITY
AGGREGATE STATISTICS FOR BETA 1

Independence				
Estimator	Mean	Standard Error	Minimum	Maximum
PCM	1.76	0.38	0.25	8.32
PEM	1.91	0.64	-1.80	8.66
ADHOC	2.12	0.68	-1.13	8.84
I/M	2.827	1.15	-2.82	12.85
True Beta	2.833			
OLS	2.99	1.19	-2.92	13.21
Dependence				
Estimator	Mean	Standard Error	Minimum	Maximum
ADHOC	2.582	0.54	-0.15	6.48
PCM	2.584	0.47	0.71	4.06
PEM	2.59	0.52	-0.08	9.04
True BETA	2.608			
I/M	2.79	1.05	-1.44	9.01
OLS	2.92	1.09	-1.49	9.39

Qualitatively, the results were the same for β_2 and β_3[14]. So, under the assumption of independence, INDIV/MATRIX is a better estimator than OLS both by accuracy and distribution. Relaxing that assumption shows that all of the methods are accurate with PEM, PCM, and ADHOC being favored both on accuracy and error distribution.

[14]See Appendix A.4, Tables A4.1 and A4.2

6.2.4 High Multicollinearity

Another pair of simulations on the independence/dependence assumption was performed under a high multicollinearity scenario (HMI and HMD) by altering the distribution of the trace of eigenvalue matrix in a similar fashion as was done with the Gorman and Toman data.

Redistributed Eigenvalues		
λ_1	λ_2	λ_3
1000	8	0.1782

Focusing on the relative performance of the estimator set under high multicollinearity conditions in Table 6.26, PEM, PCM, and ADHOC are clearly better then INDIV/MATRIX or OLS with or without independence. In both instances, the three have much smaller MSE's and errors. The two most striking outcomes are the similarity of the INDIV/MATRIX estimates to OLS and the very small maximum value for ADHOC.

TABLE 6.26
AGGREGATE STATISTICS
FOR THE PSEUDORANDOM DATASET HMI AND HMD

Independence				
Estimator	MSE	Standard Error	Minimum	Maximum
PEM	1.49	0.85	0.08	12.66
PCM	1.55	0.96	0.05	12.66
ADHOC	1.59	0.86	0.07	6.20
I/M	1.757	1.04	0.07	12.65
OLS	1.758	1.04	0.07	12.66
Dependence				
Estimator	MSE	Standard Error	Minimum	Maximum
PEM	1.14	0.64	0.02	9.92
PCM	1.27	0.80	0.09	10.03
ADHOC	1.37	0.79	0.07	5.00
I/M	1.59	0.91	0.12	10.06
OLS	1.59	0.91	0.11	10.06

Comparing LMI, LMD, HMI, and HMD, the only major change is the improved performance of ADHOC in a high multicollinearity situation. The other four have roughly the same MSE and standard errors under both assumptions and their relative ranking is the same as well. However, the ADHOC MSE dropped about 20% in both cases while the errors actually rose; hence, the amount of bias was reduced. Based on these simulations, the unbounded-risk estimators PEM and PCM are the best performers, ADHOC is consistent if small standard errors are the goal, and INDIV/MATRIX and OLS are consistent by both criteria with INDIV/MATRIX having slightly lower MSE and standard errors in the low multicollinearity case.

Table 6.27 lists the results for β_1. Under independence, INDIV/MATRIX and OLS obtained the closest estimates with ADHOC being closer than PCM and PEM. With dependence, all estimators overstated True BETA, but the relative performance switched from the independence situation because PCM. PEM, and ADHOC were closer than INDIV/MATRIX and OLS.

In both scenarios, PCM, PEM, and ADHOC had consistently smaller standard errors than INDIV/MATRIX and OLS.

In these simulations, ADHOC could be deemed the best choice since it had very small errors while it was able to estimate True BETA relatively precisely

TABLE 6.27
HIGH MULTICOLLINEARITY
AGGREGATE STATISTICS FOR BETA 1

Independence				
Estimator	Mean	Standard Error	Minimum	Maximum
PCM	2.15	0.84	-0.84	13.01
PEM	2.29	0.93	-1.80	13.21
ADHOC	2.41	0.81	-0.76	11.78
True BETA	2.83			
I/M	2.990	1.20	-2.92	13.21
OLS	2.992	1.20	-2.92	13.21
Dependence				
Estimator	Mean	Standard Error	Minimum	Maximum
True BETA	2.608			
PCM	2.63	0.55	0.24	9.04
PEM	2.64	0.62	-0.31	9.39
ADHOC	2.72	0.59	0.07	7.70
I/M	2.92	1.09	-1.49	9.38
OLS	2.92	1.09	-1.49	9.39

The results for β_2 and β_3[15] were slightly different. Under independence INDIV/MATRIX and OLS estimated True BETA well but still had large errors; ADHOC was closer than PEM or PCM. Under dependence, ADHOC was slightly more accurate than PEM or PCM whereas INDIV/MATRIX and OLS became the most accurate estimators of True BETA.

[15]See Appendix A.4, Tables A.4.3 and A.4.4.

Comparing and summarizing both multicollinearity situations, there does not appear to be an overall 'winner'. PEM, PCM, and ADHOC produce consistently small standard errors, but there is variability in their capability to estimate the parameter precisely.

CHAPTER 7

CONCLUSION AND ADDITIONAL TOPICS

7.1 Conclusion

The most significant contribution of this dissertation is that a very general, unifying regression structure called principal elements regression was introduced. The manner in which it embraces such traditional regression techniques as restricted least squares, principal components, and factor analysis was developed analytically. In addition, it was also demonstrated that PER includes the relatively new fractional approach of Lee as a special case. Any method that imposes a weighting scheme on the individual elements of the eigenvector matrix is a member of the PER class of estimators.

Second, it was also shown that PER forms other than PCR can produce a variance less than or equal to PCR with an equal number of eigenvector elements zeroed or deleted. Further, new PER forms entitled ADHOC, INDIV, and MATRIX quite possibly can produce a lower variance with fewer eigenvector elements modulated (subrestrictions imposed). Indeed, these new estimators do not require the harsh constraint of zeroing or deleting principal elements as does PCR, but can selectively shrink individual elements.

The third contribution demonstrated the inadequacy of the traditional comparative statistics of linear restrictions and rank and subsequently defined the more informative terms modulation, subrestriction, and mode. The two traditional notions, while being useful for some biased regression techniques, actually conceal an estimators' impact on the elements of the eigenvector matrix. On the other hand, modulation accounts for differential treatment of the individual elements--subrestriction reflects the matrix-wide constraints more accurately than does restriction and the mode quantifies the number of unrestricted elements. A single restriction on a column of a matrix having k elements equals 1 regardless of how many elements are retained--$(0 or 1 or \ldots or\ k)$--resulting in a rank of $(k - 1)$, whereas the mode of the matrix could range from $\left(k^2 - k\right)$ to k^2.

Since a sample usually suffers from ill-conditioning due to multicollinearity, and since multicollinearity is intractable, attention is diverted to loss functions such as mean squared error. Thus, a subtle shift in focus from sample conditions to population characteristics occurs. Monte carlo experiments are an attempt to determine the population characteristics and not explicitly eliminate the multicollinearity in the sample data. The experimental portion of this work utilized monte carlo simulations on two datasets to explore the empirical risks of an estimator set.

The new PER forms ADHOC, INDIV, and MATRIX are a category of shrinkage estimators. So, they were evaluated in high multicollinearity experiments as part of an estimator set that included a positive-part Stein-rule and a truncated Stein-rule that were conditioned with each of the six base estimators OLS, PEM, PCM, ADHOC, INDIV, and MATRIX. Those

simulations on the Gorman and Toman dataset showed a clear tendency for ADHOC, INDIV, and MATRIX to produce relatively low mean squared error coincident with relatively low standard errors. As well, these three PER forms produced impressive results when the components of the beta vector were inspected--there was a tendency toward smaller standard errors for the individual parameter estimates which meant they were able to locate a larger amount of the probability distribution closer to the true coefficient.

Low multicollinearity conditions were simulated on the Gorman and Toman data by dividing the trace of the eigenvalue matrix equally among the ten eigenvalues. Predictably, subsequent analysis demonstrated, with one exception, that the estimators were unable to reduce mean squared error below the OLS amount. The exception, ADHOC, produced an exceptionally low MSE.

Uncertainty always lingers even from thousands of monte carlo outcomes derived from a sample. This dissertation circumvented that limitation by declaring a pseudorandom dataset to be a population and proceeded with an analysis of the relative performance of six estimators, OLS, PCM, PEM, ADHOC, INDIV, and MATRIX. It would be too much to expect for one estimator to dominate the others so general conclusions of superiority are elusive.

With the UNKIND data where certain assumptions were violated, ADHOC and INDIV/MATRIX had the smallest MSE's and standard errors. In low multicollinearity conditions, ADHOC consistently had small errors but the largest MSE whether the explanatory variables and residuals were independent or dependent. INDIV/MATRIX was slightly better than OLS in

terms of both MSE and errors while the principal elements Mundlak form was the best overall. When high multicollinearity was imposed on the data, the MSE of ADHOC improved considerably while retaining low standard errors. On the other hand, INDIV/MATRIX changed little compared to OLS while PEM retained its superiority.

The small dimensions of the pseudorandom data allowed a more thorough examination of individual parameter estimates than did the Gorman and Toman data. ADHOC and INDIV/MATRIX parameter estimates of the problematic UNKIND data were the most accurate for β_1 and degenerated for β_2 and β_3. Yet, the standard errors remained consistently small. It is worth noting that OLS was a relatively inaccurate estimator of all β_i and had consistently large errors as well. The simulations with multicollinearity scenarios produced ambiguous results. With independence INDIV/MATRIX and OLS were the best estimators of all β_i and ADHOC was slightly less accurate. With dependence, there was less distinction among the three. Again, ADHOC had consistently small standard errors while INDIV/MATRIX was similar to OLS.

Despite the disparity of results, in all simulations on both datasets, a feature of ADHOC was small standard errors, a feature of OLS was large errors, and statistics on the other estimators were discordant.

7.2 Additional Topics

> In the theory and practice of econometrics the model, the method and the data are all interdependent links in the research chain. Seldom, however, are the economic and statistical models correctly specified, the data free of measurement or specification errors, the methods or decision rules used in estimation "optimal" and the inferences free of distortions. Judge and Bock (1978)

The practice of using monte carlo simulations allows pre-testing of estimators so that the researcher can make an informed choice about which estimation technique appears to work best in a particular scenario. Computer technology has evolved to the state where simulations can be performed on desktop models that would have strained mainframes as few as ten or fifteen years ago. Indeed, this environment presents opportunities for econometricians to be creative in the presence of obstacles like scarce data, poorly conditioned data, and small samples.

Judge and Yancey (1986) on inequality estimation: "What we did not know then, and what took over a decade to sort out, was how to write the estimator so that the underlying sampling properties could be determined." There are two points; one, progress is not always instantaneous because the most efficient way to approach a problem may be elusive and two, a new idea like principal elements regression that is so broad must be explored on an indeterminable number of fronts. In the future, the principal elements framework must be examined on the same grounds as other regression techniques introduced over past years (for example, the theoretical properties).

Heretofore, research has focused on reducing the dimensions of the eigenvector matrix in an effort to mitigate the effects of multicollinearity. If the sum of the information in the matrix was normalized to 1, then the usual approach can be thought of as reducing that value to less than 1. But the problem may not be the amount of information. The confounding condition called multicollinearity is a structural problem present in the sample, so perhaps a more fruitful approach would be to retain all of the information and merely reallocate it, being guided by the information in the principal elements. Recall that prediction along the plane of collinearity can be beneficial so it might be feasible to reorient the data along that plane.

Even the quadratic loss function as the objective function is suspect. It is rather benign and ubiquitous--perhaps because of its widespread popularity its appropriateness is seldom questioned.

An illustration:

Suppose that a researcher is developing an estimation technique to guide a firm that wishes to maximize profits. Suppose the price relationship is

$$p = \beta_0 + \beta_1 Q ,$$

its estimate is

$$\hat{p} = \hat{\beta}_0 + \hat{\beta}_1 Q ,$$

the profit function is

$$\begin{aligned} \Pi &= TR - TC \\ &= pQ - f(Q), \end{aligned}$$

and its estimate is

$$\hat{\Pi} = \hat{\beta}_0 Q + \hat{\beta}_1 Q^2 - f(Q).$$

The true loss to the firm is the difference between desired profits and estimated profits

$$Loss = \Pi - \hat{\Pi}$$

and the risk is expected loss

$$\begin{aligned} Risk &= E\left(\Pi - \hat{\Pi}\right) \\ &= E\left(\beta_0 - \hat{\beta}_0\right)Q + E\left(\beta_1 - \hat{\beta}_1\right)Q^2 - E(f(Q)) + \hat{f}(Q) \ . \end{aligned}$$

Compare that to mean squared error loss

$$MSE = E\left(\beta_0 - \hat{\beta}_0\right)^2 + E\left(\beta_1 - \hat{\beta}_1\right)^2 .$$

In this example, the correct criterion for estimator selection should be the expected loss in profits and not the expected prediction loss.

Pursuing this line of research does not depend on principal elements regression, but this dissertation has drawn attention to the correct design of all aspects of an experiment. Depending upon the evaluation criterion, an estimator that performs favorably in one case may not be as desirable under a different standard. With the computing power currently available, it is compelling for a researcher to use monte carlo evaluation techniques to evaluate several competing estimators as well as experiment with different loss functions.

APPENDIX

A.1 The Ordering Matrix $\overset{o}{A}$

The following two tables list the principal elements $\left(a_{ij}^2/\lambda_j\right)$ first and then the ordering matrix $\overset{o}{A}$ that reflects the relative magnitude of each element of the Gorman and Toman data.

Principal Elements Matrix

8.74E-9	1E-8	0.00001	0.00068	0.00296	0.00241	5.68E-7	0.0145	0.00009	0.00056
7.65E-7	4.9E-7	0.00023	0.00086	0.00019	0.00015	0.00006	0.00014	0.00002	2.02E-6
372E-13	454E-13	2.82E-8	9.27E-7	924E-14	1.82E-6	0.00024	0.01576	6.24652	0.08711
373E-12	23E-11	1.22E-8	5.22E-6	2.4E-6	5.05E-7	0.17021	0.00748	0.00306	0.00041
518E-14	811E-14	1.19E-8	5.96E-8	1.1E-7	5.36E-7	9.22E-6	0.00197	0.0018	145.193
333E-12	182E-12	9.07E-7	0.00001	0.00005	0.00006	0.00159	0.7122	0.14346	0.35482
1.03E-6	7.42E-7	0.00014	0.00058	0.00023	0.00048	0.00022	0.00019	0.00074	0.00053
2.38E-9	1.23E-8	0.00003	0.00006	0.00257	0.00537	0.00009	0.00025	0.00089	0.0011
7.77E-8	0.00003	1.63E-8	3.09E-7	3.6E-7	5.44E-7	9.64E-8	2.55E-7	9.23E-9	7.73E-6
6.07E-9	942E-12	0.00063	0.00056	0.00014	0.00171	0.00004	0.00018	0.00016	0.00003

The Ordering Matrix $\overset{o}{A}$

88	86	57	24	13	15	69	[9]	45	28
67	73	34	22	37	41	49	44	55	62
97	96	81	65	98	63	33	[8]	[2]	[7]
92	94	84	60	61	72	[5]	[10]	12	31
100	99	85	80	77	71	58	16	17	[1]
93	95	66	56	50	48	19	[3]	[6]	[4]
64	68	43	26	35	30	36	38	23	29
90	83	52	47	14	11	46	32	21	20
79	53	82	75	74	70	78	76	87	59
89	91	25	27	42	18	51	39	40	54

This table emphasizes the way the principal elements regression form can exploit its flexibility compared to principal elements regression. The principal elements estimator that zeroed the last column would only capture three of the smallest variance elements. A PER form that also selected ten elements would focus on those that are emphasized.

A.2 Eigenvector Matrices

The following are examples of eigenvector matrices. The 3x3 matrices are from a single iteration on the pseudorandom data and are presented first because their small dimension makes them easier to examine than those from the Gorman and Toman data that follow.

Eigenvectors of X`X

0.64316	-0.7657	-0.0092
0.57076	0.48741	-0.6608
0.51047	0.41973	0.7505

Eigenvectors of ADHOC

0.62141	-0.1107	-0.0001
0.55147	0.07048	-0.0083
0.49321	0.06069	0.00948

Eigenvectors of INDIV/MATRIX

0.81387	-0.0181	-0.0038
0.48393	-0.0108	-0.0023
0.31467	-0.007	-0.0015

Eigenvectors of PCM

0.64316	0	0
0.57076	0	0
0.51047	0	0

Eigenvectors of PEM

0.64316	0	-0.0092
0.57076	0	0
0.51047	0	0

The following tables are examples of the eigenvector matrices from the same iteration on the Gorman and Toman data. Note that PCM zeroed 9 components (90 elements) while PEM only zeroed 10 elements.

Eigenvectors of X`X

0.06797	-0.0182	-0.1067	0.49678	0.69368	0.48621	0.00181	0.13882	0.00365	0.00196
0.63575	-0.127	-0.4723	-0.5566	0.17728	0.12229	0.01788	-0.0136	0.00184	0.00012
0.00444	-0.0012	-0.0052	0.01832	0.00004	0.01337	-0.0372	-0.1447	0.98819	0.02446
0.01403	-0.0027	0.0034	0.04347	-0.0197	0.00704	0.9935	-0.0997	0.02186	-0.0017
0.00165	-0.0005	0.00336	-0.0046	-0.0042	-0.0073	0.00731	0.05111	-0.0168	0.99847
0.01326	-0.0024	0.02938	-0.0737	-0.0869	-0.0753	0.09598	0.97275	0.14976	-0.0494
0.73843	-0.1562	0.36582	0.45868	-0.1931	-0.2172	-0.0358	-0.0157	-0.0107	-0.0019
0.03547	-0.0201	-0.1736	0.1447	-0.647	0.7264	-0.0222	0.01826	-0.0118	0.00275
0.20262	0.97913	-0.0039	0.01058	-0.0076	0.00731	-0.0007	0.00058	0.00004	0.00023
0.05665	-0.0056	0.77505	-0.452	0.15314	0.40964	0.0148	-0.0153	0.00496	-0.0004

Eigenvectors of ADHOC

0.04619	-0.0023	-0.0008	0.01784	0.2332	0.0091	0.00033	0.00819	0.00001	0.00021
0.43206	-0.0159	-0.0036	-0.02	0.0596	0.00229	0.00329	-0.0008	6.23E-6	0.00001
0.00301	-0.0002	-399E-7	0.00066	0.00001	0.00025	-0.0068	-0.0085	0.00335	0.00266
0.00954	-0.0003	0.00003	0.00156	-0.0066	0.00013	0.18279	-0.0059	0.00007	-0.0002
0.00112	-0.0001	0.00003	-0.0002	-0.0014	-0.0001	0.00135	0.00301	-567E-7	0.10857
0.00901	-0.0003	0.00023	-0.0026	-0.0292	-0.0014	0.01766	0.05736	0.00051	-0.0054
0.50184	-0.0196	0.00281	0.01647	-0.0649	-0.0041	-0.0066	-0.0009	-364E-7	-0.0002
0.02411	-0.0025	-0.0013	0.0052	-0.2175	0.0136	-0.0041	0.00108	-4E-5	0.0003
0.1377	0.12268	-302E-7	0.00038	-0.0026	0.00014	-0.0001	0.00003	1.29E-7	0.00003
0.0385	-0.0007	0.00596	-0.0162	0.05148	0.00767	0.00272	-0.0009	0.00002	-465E-7

Eigenvectors of INDIV/MATRIX

19.9934	-1.2965	-0.0513	-0.0694	0.17142	0.02585	0.0216	0.00545	0.00044	0.00055
65.34702	-0.3467	-0.0137	-0.0186	0.04584	0.00691	0.00578	0.00146	0.00012	0.00015
52.339	-3.3939	-0.1342	-0.1816	0.44874	0.06767	0.05653	0.01426	0.00114	0.00143
42.1134	-2.7309	-0.108	-0.1461	0.36107	0.05445	0.04549	0.01148	0.00092	0.00115
1051.19	-68.165	-2.6958	-3.6469	9.01255	1.35917	1.13541	0.2865	0.02288	0.02874
9.71739	-0.6301	-0.0249	-0.0337	0.08331	0.01256	0.0105	0.00265	0.00021	0.00027
-9.4205	0.61088	0.02416	0.03268	-0.0808	-0.0122	-0.0102	-0.0026	-0.0002	-0.0003
-4.3305	0.28081	0.01111	0.01502	-0.0371	-0.0056	-0.0047	-0.0012	-0.0001	-0.0001
-0.2747	0.01781	0.0007	0.00095	-0.0024	-0.0004	-0.0003	-0.0001	-598E-8	-751E-8
3.80518	-0.2467	-0.0098	-0.0132	0.03262	0.00492	0.00411	0.00104	0.00008	0.0001

Eigenvectors of PCM

0.06797	0	0	0	0	0	0	0	0	0
0.63575	0	0	0	0	0	0	0	0	0
0.00444	0	0	0	0	0	0	0	0	0
0.01403	0	0	0	0	0	0	0	0	0
0.00165	0	0	0	0	0	0	0	0	0
0.01326	0	0	0	0	0	0	0	0	0
0.73843	0	0	0	0	0	0	0	0	0
0.03547	0	0	0	0	0	0	0	0	0
0.20262	0	0	0	0	0	0	0	0	0
0.05665	0	0	0	0	0	0	0	0	0

Eigenvectors of PEM

0.06797	-0.0182	-0.1067	0	0.69368	0	0.00181	0.13882	0.00365	0.00196
0.63575	-0.127	0	0	0.17728	0.12229	0.01788	-0.0136	0.00184	0.00012
0.00444	-0.0012	-0.0052	0.01832	0.00004	0.01337	-0.0372	-0.1447	0	0.02446
0.01403	-0.0027	0.0034	0.04347	-0.0197	0.00704	0.9935	-0.0997	0.02186	-0.0017
0.00165	-0.0005	0.00336	-0.0046	-0.0042	-0.0073	0.00731	0.05111	-0.0168	0
0.01326	-0.0024	0.02938	-0.0737	-0.0869	-0.0753	0.09598	0	0	-0.0494
0.73843	-0.1562	0	0.45868	-0.1931	-0.2172	0.0358	-0.0157	-0.0107	-0.0019
0.03547	-0.0201	-0.1736	0.1447	-0.647	0	0.0222	0.01826	-0.0118	0.00275
0.20262	0	-0.0039	0.01058	-0.0076	0.00731	0.0007	0.00058	0.00004	0.00023
0.05665	-0.0056	0	-0.452	0.15314	0	0.0148	-0.0153	0.00496	-0.0004

A.3 Gorman and Toman Data

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	Y
7.89	79	0.57	2.50	0.19	1.422	91.7	5.00	26	4.1	66
8.68	83	0.50	1.52	0.19	1.642	87.8	4.49	7	2.7	120
5.61	82	0.68	1.69	0.22	2.310	87.0	8.46	26	1.6	293
14.75	78	0.51	1.21	0.18	1.150	87.0	1.77	6	4.3	35
7.26	86	0.51	1.45	0.20	1.661	88.4	5.53	13	2.6	160
6.62	82	0.49	1.22	0.20	1.799	87.6	3.61	3	0.1	106
6.25	88	0.50	1.20	0.23	1.892	88.9	4.16	6	3.7	104
14.55	77	0.55	2.91	0.17	1.031	90.2	2.86	3	3.5	37
9.44	74	0.50	1.84	0.20	1.348	88.8	5.13	10	5.3	84
6.54	84	0.53	1.43	0.22	1.638	91.8	3.49	10	3.9	132
8.39	75	0.63	1.45	0.19	1.305	92.2	3.45	11	4.4	71
9.68	72	0.60	1.59	0.19	1.335	90.2	3.41	12	4.1	123
10.09	76	0.62	1.47	0.19	1.180	90.5	4.14	8	6.3	67
8.60	82	0.54	1.64	0.20	1.370	89.9	7.80	26	6.6	141
11.48	77	0.65	1.93	0.18	1.031	91.4	4.65	28	6.6	77
8.84	73	0.59	1.58	0.20	1.347	91.4	5.03	13	11.9	125
12.06	72	0.78	1.97	0.09	0.641	83.2	9.31	4	0.2	52
12.21	69	0.84	1.31	0.12	0.618	85.3	6.69	35	0.6	25
6.72	80	0.51	2.87	0.22	1.731	87.5	5.12	10	5.8	102
8.65	79	0.54	1.46	0.16	1.381	89.8	7.97	75	3.8	206
8.01	76	0.51	1.92	0.17	1.466	88.8	3.65	19	5.4	190
5.96	70	0.48	1.65	0.23	1.994	90.2	4.17	16	7.6	270
5.86	71	0.43	1.96	0.23	2.131	89.1	5.49	6	6.0	390
5.63	71	0.43	1.51	0.22	2.204	88.8	4.29	30	10.4	458
6.19	71	0.44	1.62	0.26	2.101	90.3	1.53	17	10.7	129
5.90	72	0.42	2.05	0.24	2.063	87.6	7.66	12	9.4	268
6.57	74	0.44	1.04	0.24	1.909	88.5	2.56	9	17.9	188
5.84	73	0.42	1.78	0.22	2.017	87.2	5.48	33	10.7	310
5.84	69	0.42	1.90	0.23	2.011	86.4	2.06	22	12.9	260
8.23	72	0.46	2.02	0.17	1.589	89.4	1.83	50	12.5	190
8.12	82	0.47	1.76	0.19	1.706	90.1	0.65	185	6.5	164
9.25	83	0.50	1.31	0.19	1.674	94.9	3.12	28	11.7	138
8.31	82	0.47	1.34	0.20	1.667	90.3	5.13	4	9.6	202
7.80	85	0.63	1.86	0.22	1.962	92.2	3.16	24	8.5	102
6.58	77	0.62	1.51	0.24	1.959	95.0	1.33	15	16.6	160
8.97	76	0.62	1.92	0.24	1.592	95.5	2.13	22	8.8	101

A.4 Pseudorandom Data

The following SAS/IML code generated the pseudorandom data:

```
cx = 1;  cb = 3;  cr = 9; n = 10;    t = 1000;    k = 3;
seedx = 3 ;  seedb = 7 ;  seedr = 11;
xdimen = j(t,k,seedx);
   xpop = cx * uniform(xdimen);
bdimen = j(k,1,seedb);
   beta = cb * uniform(bdimen);
rdimen = j(t,1,seedr);
   resid = ((cr * normal(rdimen)) - .44772) / 9.03432;
ypop = xpop * beta + resid;
```

The following SAS/IML code randomly selected 10 observations:

```
s_size = 10;
s_obs = int((99 * uniform(s_size)))` + 1;
 r = rank(s_obs);
 term1 = s_obs;
 s_obs(|,rank(s_obs)|) = term1;
use sasdata.modldata;
read point s_obs into dupe;
y = dupe(|,1|);
x = dupe(|,{2 3 4}|);
```

A sample of the data:

Y	X1	X2	X3	residuals
5.8876	0.7589	0.8226	0.0652	0.2647
8.7216	0.9138	0.8828	0.6373	3.1423
-3.7526	0.1162	0.2477	0.9408	-4.4381
12.4245	0.9947	0.8304	0.6000	7.0442
10.7232	0.5304	0.1277	0.5414	9.9631
0.9257	0.0058	0.9048	0.5113	-4.0167
6.4072	0.9355	0.8054	0.4830	1.1213
10.8766	0.8317	0.9040	0.7212	5.3326
1.6304	0.7133	0.4079	0.8125	-0.7283
-4.4637	0.8303	0.1613	0.0561	-6.1930

TABLE A.4.1
LOW MULTICOLLINEARITY
AGGREGATE STATISTICS FOR BETA 2

Independence				
Estimator	Mean	Standard Error	Minimum	Maximum
I/M	1.32	1.06	-7.63	6.59
OLS	1.40	1.13	-7.97	6.85
True BETA	1.65			
ADHOC	1.69	0.52	2.23	5.65
PEM	1.91	0.65	-2.94	6.83
PCM	2.06	0.29	-1.86	3.14
Dependence				
Estimator	Mean	Standard Error	Minimum	Maximum
I/M	2.11	1.01	-4.92	6.50
OLS	2.201	1.05	-5.23	6.94
True BETA	2.199			
ADHOC	2.33	0.50	-1.22	4.97
PEM	2.50	0.49	-1.29	5.22
PCM	2.52	0.42	0.72	4.06

TABLE A.4.2
LOW MULTICOLLINEARITY
AGGREGATE STATISTICS FOR BETA 3

Independence				
Estimator	Mean	Standard Error	Minimum	Maximum
True BETA	1.10			
I/M	1.31	1.07	-4.40	7.47
OLS	1.39	1.14	-4.52	7.97
ADHOC	1.50	0.54	-2.52	5.63
PEM	1.67	0.63	-3.10	7.97
PCM	1.72	0.33	-0.24	7.97
Dependence				
Estimator	Mean	Standard Error	Minimum	Maximum
ADHOC	2.34	0.48	-1.54	5.59
PCM	2.39	0.47	0.51	3.93
PEM	2.40	0.53	-3.80	5.59
I/M	2.50	0.90	-4.55	6.38
OLS	2.62	0.94	-4.65	6.71
True BETA	2.92			

TABLE A.4.3
HIGH MULTICOLLINEARITY
AGGREGATE STATISTICS FOR BETA 2

Independence				
Estimator	Mean	Standard Error	Minimum	Maximum
I/M	1.404	1.13	-7.96	6.85
OLS	1.405	1.13	-7.97	6.85
True BETA	1.65			
ADHOC	1.74	062	-5.74	4.98
PEM	1.78	0.73	-3.11	5.49
PCM	1.88	0.60	-2.67	4.34
Dependence				
Estimator	Mean	Standard Error	Minimum	Maximum
True BETA	2.199			
I/M	2.20	1.05	-5.23	6.94
OLS	2.20	1.05	-5.23	6.93
ADHOC	2.41	0.57	-2.08	5.14
PEM	2.43	0.59	-2.76	5.42
PCM	2.49	0.52	-0.52	5.42

TABLE A.4.4
HIGH MULTICOLLINEARITY
AGGREGATE STATISTICS FOR BETA 3

Independence				
Estimator	Mean	Standard Error	Minimum	Maximum
True BETA	1.10			
I/M	1.39	1.14	-4.52	7.96
OLS	1.39	1.14	-4.52	7.97
ADHOC	1.50	0.70	-3.91	6.25
PEM	1.51	0.80	-5.05	7.97
PCM	1.58	0.70	-5.05	7.97
Dependence				
Estimator	Mean	Standard Error	Minimum	Maximum
PCM	2.42	0.55	-3.80	5.48
PEM	2.43	0.61	-3.32	5.49
ADHOC	2.47	0.56	-3.05	5.39
I/M	2.61	0.94	-4.65	6.71
OLS	2.62	0.94	-4.65	6.71
True BETA	2.92			

BIBLIOGRAPHY

Adkins, L. C., 1991, Prior information in regression: To choose or not to choose?, paper presented to the Southern Economic Association, Memphis, November.

Baranachik, A. J., 1964, Multiple regression and estimation of the mean of a multivariate normal distribution, Technical Report No. 51, Department of Statistics, Stanford University, Stanford, CA.

Brunson, K. D., L. C. Marsh, and T. Urbine, 1990, Principal elements regression, paper presented to the Midwest Economics Association, Chicago, March.

Brunson, K. D. and L. C. Marsh, 1991a, Using SAS/IML® software in a new approach to dealing with multicollinearity, paper presented at the SUGI 16 Proceedings, New Orleans, February.

Brunson, K. D. and L. C. Marsh, 1991b, Mean square error reduction through improved data analysis transformation: A new principal elements statistical method, paper presented to the Eastern Economics Association, Pittsburgh, March.

Brunson, K. D. and L. C. Marsh, 1991c, A principal elements approach to forecasting fatalities in U. S. bituminous coal production, paper presented to the Third Interamerican Applied Statistics Seminar, Mexico City, September.

Brunson, K. D. and L. C. Marsh, 1991d, Comparing alternative shrinkage estimators to reduce mean square error, paper presented to the Southern Economic Association, November.

Chen, C. W., 1974, An optimal property of principal components, Communications in Statistics 3 (10), 979-983.

Dey, D. K. and J. O. Berger, 1983, On truncation of shrinkage estimators in simultaneous estimation of normal means, Journal of the American Statistical Association 78 (384), 865-869.

Dhrymes, P. J., 1984, Mathematics for econometrics. (Springer-Verlag, New York).

Donnell, D. J., 1987, Additive principal components: A method for estimating additive constraints with small variance from multivariate data. (Ph.D. dissertation, University of Washington).

Efron, B. and C. Morris, 1973, Stein's estimation rule and its competitors -- an empirical Bayes approach, Journal of the American Statistical Association 68 (341), 117-130.

Fomby, T. B. and R. C. Hill, 1988, Small-sigma approximations and the minimaxity of Stein-rules under nonnormality, Advances in Econometrics 7, 193-209.

Fomby, T.B. and R. C. Hill, 1989, The geometry of least squares prediction, working paper, Louisiana State University, Baton Rouge, LA 70803.

Fomby, T. B., R. C. Hill, and S. R. Johnson, 1978, An optimal property of principal components in the context of restricted least squares, Journal of the American Statistical Association 73, 191-193.

Gabriel, K. R., 1981, Interpreting multivariate data, edited by Vic Barnett. (Wiley, Chichester).

George, E. I., 1986, Minimax multiple shrinkage estimation, Annals of Statistics 14, 188-205.

Gorman, J. W. and R. J. Toman, 1966, Selection of variables for fitting equations to data, Technometrics 8, 27-51.

Green, E. J. and W. E. Strawderman, 1991, A James-Stein type estimator for combining unbiased and possibly biased estimators, Journal of the American Statistical Association 86 (416), 1001-1006.

Greenberg, E., 1975, Minimum variance properties of principal component regression, Journal of the American Statistical Association 70, 194-197.

Gunst, T. F. and R. L. Mason, 1976, Generalized mean squared error properties of regression estimators, Communications in Statistics A(15), 1501-1508.

Gunst, T. F. and R. L. Mason, 1977, Biased estimation in regression: An evaluation using mean squared error, Journal of the American Statistical Association 72 (359), 616-628.

Hill, R. C., and G. Judge, 1987, Improved prediction in the presence of multicollinearity, Journal of Econometrics 35, 83-100.

Hoerl A. E., and Kennard R. W., 1970, Ridge regression: Biased estimation for nonorthogonal problems, Technometrics 12, 55-67.

Jackson, J. E., 1991, A user's guide to principal components. (Wiley, New York).

James, W. and C. Stein, 1961, Estimation with quadratic loss, Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, University of California Press, Berkeley, CA, 316-379.

Judge, G. G. and M. E. Bock, 1978, The statistical implications of pre-test and Stein-rule estimators in econometrics. (North-Holland, Amsterdam).

Judge, G. G. and T. A. Yancey, 1986, Improved methods of inference in economics. (North-Holland, Amsterdam).

Kendall, M. G., 1957, A course in multivariate analysis. (London, Griffin).

Kmenta, Jan, 1986, Elements of econometrics. (Macmillan, New York).

Lee, W. W. and J. B. Birch, 1986, Fractional principal components regression: A general approach to biased estimation, paper presented to the Physical and Engineering Sciences Section of the American Statistical Association, Chicago, August 18.

Lee, W. W., 1986, Fractional principal components regression: A general approach to biased estimators. (Ph.D. dissertation, Virginia Polytechnic Institute).

Magnus, J. R. and H. Neudecker, 1988, Matrix differential calculus with applications in statistics and econometrics. (Wiley, Chichester).

Mansfield, E. R., 1975, Principal component approach to handling multicollinearity in regression analysis. (Ph.D dissertation, Southern Methodist University).

Marquardt, D. W., 1970, Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation, Technometrics 12, 591-612.

Marsh, L. C. and M. J. Radzicki, 1986, Minimum mean square error of a generalized Stein estimator, Economics Letters 21, 333-335.

Mason, R. L., R. F. Gunst , and J. T. Webster, 1975, Regression analysis and problems of multicollinearity, Communications in Statistics 4 (3), 277-292.

Massy, W. F., 1965, Principal components regression in exploratory statistical research, Journal of the American Statistical Association 60, 234-256.

Miller, A. J., 1990, Subset selection in regression. (Chapman and Hall, London).

Mundlak, Y., 1981, On the concept of nonsignificant functions with its implications for regression analysis, Journal of Econometrics 16, 139-150.

Oksanen, E. H., 1988, Principal components in econometrics, Communications in Statistics A 17 (8), 2507-2532.

Park, S. H., 1981, Collinearity and optimal restrictions on regression parameters for estimating responses, Technometrics 23 (3), 289-295.

Rogers, Gerald S., 1980, Matrix derivatives. (Marcel-Dekker, New York).

Stein, C., 1981, Estimation of the mean of a multivariate normal distribution, Journal of Econometrics 44, 189-213.