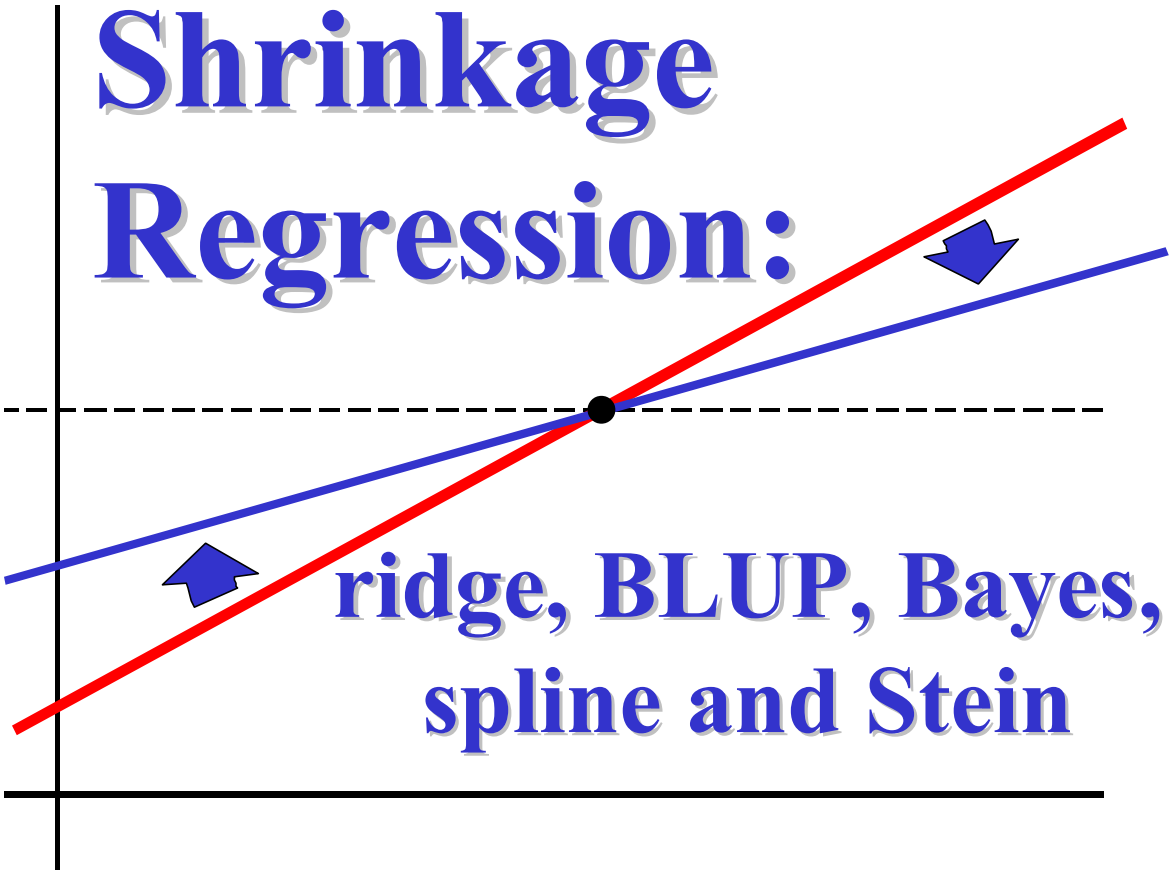
Shrinkage Regression:



Chapter 00: Preface and Table of Contents

Bob Obenchain, Ph.D.
softR$_X$ freeware
13212 Griffin Run
Carmel, Indiana 46033-8835

To the great ladies who shaped and enriched my life...

Lynne, Tiffany, Anne, Lottie and mistress mathematics.

Preface

This book introduces, motivates, and explores a variety of methodologies for “shrinking" the regression coefficients that result when fitting models to possibly ill-conditioned and/or imprecise (errors-in-variables) data. The four main, broad categories of shrinkage methodology we consider here are ridge, BLUP, Bayes and Stein estimation. In ridge estimation, the underlying model usually views the unknown, true regression coefficients as being **fixed** constants. And the corresponding objective for shrinkage is to exploit variance-bias trade-offs to reduce mean-squared-error in estimation. In this same context, Stein methods are highly specialized forms of **uniform** coefficient shrinkage that focus interest on admissibility and/or minimax properties relative to specific, **univariate** (scalar valued) measures of risk. In BLUP and Bayes estimation, true regression coefficients are viewed as being unknown realizations of **random variables**, but estimation procedures suggested by these approaches again correspond to various forms of coefficient shrinkage.

Part One: Shrinkage Regression Theory and Methodology

We place special emphasis in Part One of our exposition upon normal-distribution-theory, maximum likelihood formulations that unify and contrast the ridge, BLUP, and empirical Bayes approaches to shrinkage. But we also explore a spectrum of alternative motivations for classical (fixed coefficient), random coefficient, and Bayes (added information) methods. We employ a uniform collection of terminology and notation throughout, and we exploit this formalism to give insights into interrelationships among diverse methods.

When I am voicing my own personal opinions about the relative strengths and weaknesses in a methodology, I will attempt to consistently use first person, singular pronouns (I, my).

People who know me well say that they “can almost hear me talking" when they read what I write. I suspect this is partly due to my use of somewhat non-standard typography. For example, I tend to place words in **bold face** or ALL CAPS for emphasis. And I tend to place quotation marks (“") around words or phrases that I am using in a somewhat ambiguous way...as in an analogy that is much less than complete. If I have trouble choosing between a pair of words in a particular phrase, I tend to write down both words with only a slash (/) between them. I know that referees and editors have found my typography disconcerting...but, then, I don't usually give the above “explanation" when I submit a paper for publication! Anyway, if you don't find my usage of **bolds**, CAPS, “", and /s helpful, ...please try to ignore them.

Part Two: Shrinkage Regression Applications and Implementations

In the sense that practical applications of shrinkage regression methodology **can** be highly graphical, then they **should** be highly graphical ...RIGHT? (Let's hope we can all agree on at least THAT!) While the technical materials of Part One provide, hopefully, a firm foundation for shrinkage theory and methodology, fundamental questions about general regression strategies and tactics are not addressed in Part One. Rather, questions like "Which displays should I examine?" and "How should I interpret them?" and overview materials on the **psychology of graphical perception** are primary topics for Part Two of our exposition. For example, our discussion of shrinkage TRACE displays in Chapter 11 discusses numerous advantages associated with using Multicollinearity Allowance (MCAL) scaling along the horizontal (shrinkage extent) axis on ridge TRACE displays.

Because shrinkage regression techniques tend to be computationally and graphically intensive, effective study/application of the ideas outlined in this book **requires** access to state-of-the-art computer software and hardware. I have developed prototype systems for IBM-compatible (MS-DOS and Windows) personal computers for this express purpose, and I market them as freeware to maximize their availability to potential practitioners of shrinkage regression. But I also illustrate (i) computational procedures for LISP-STAT, GAUSS, SAS/IML and S as well as (ii) commercial software systems such as SAS proc MIXED, BMDP (especially 4R and 5V), and GLMM.

I hope users will find that my personal computer software systems provide an interface that is sufficiently intuitive and self-explanatory that the full depth of detailed understanding provided by this book is not a prerequisite for effective shrinkage regression applications. But, by pulling together diverse materials - ranging through introductory remarks, common misconceptions, historical commentaries, simulation results, and theoretical fine points - we provide here not only an unquestionably strong foundation but also a full set of practically useful road-maps for the theory and application of shrinkage regression.

Shrinkage Regression: ridge, BLUP, Bayes, spline & Stein

Table of Contents

CHAPTER

1. INTRODUCTION

This introductory chapter provides an short overview of our topic. We first consider the original sense in which "regression" implies "shrinkage" by tracing standard least-squares fitting terminology back to the work of Francis Galton in the late 1800's. Then we introduce our PRIMARY THEME, the **multivariate analysis point-of-view on shrinkage regression**. Next we provide a very brief thumb-nail-sketch of how modern shrinkage methods might be applied to a simple numerical example. At the end of chapter one, we detail three key motivations for exploiting the **principal axes rotation** of regressor coordinates as a basic **canonical form** for possibly ill-conditioned regressor problems.

1.1 Galton's "Shrinkage" Interpretation of Regression
1.2 The Primary "Multivariate Analysis" Theme of This Book
1.3 How are Shrinkage Regression Methods Typically Applied?
1.4 Which Part of the Book Should I Read Next?

Part One: Shrinkage Regression Theory and Methodology

2. BASIC LINEAR MODEL CONCEPTS

This chapter reviews many aspects of the theory of **general linear models** in the possibly less-than-full-rank case.

2.1 Centered Variables
2.2 The Special Case of UNCORRELATED Regressors
2.3 Canonical Form of Regressors
2.4 NUMERICAL versus STATISTICAL ILL-CONDITIONING

2.5 Eigen Decompositions
2.6 The UNCORRELATED COMPONENTS of LEAST SQUARES
2.7 STATISTICAL SIGNIFICANCE of Uncorrelated Components
2.8 Predictions, Residuals & Linear Reparameterizations that remove Ill-Conditioning
2.9 SIGNAL-to-NOISE Ratios
2.10 The Statistical Distribution of Principal Correlations
2.11 When "Should" Coefficients have "Wrong" Signs?
2.12 Tests of General Linear Hypotheses
2.13 Weighted Residual Analyses

3. SHRINKAGE REGRESSION FUNDAMENTALS

This chapter introduces generalized shrinkage (ridge regression) estimators and points out several special cases which played major roles in the early history of shrinkage regression.

3.1 Moments of Generalized Shrinkage Estimators
3.2 Shrinkage Inflation of the Residual Mean Square
3.3 The Hoerl-Kennard ORDINARY RIDGE REGRESSION Family
3.4 The TWO-PARAMETER GENERALIZED RIDGE Family
3.5 The IMPLICIT INTERCEPT Associated with Shrinkage
3.6 Shrinkage in Models Without an INTERCEPT
3.7 Shrinkage Residual Analyses

4. THE RISK OF SHRINKAGE

How much shrinkage is "best"? We start out by showing that this question is difficult to answer - even in theory when we pretend that true values of regression parameters are **known!** Then we introduce the concepts of "optimal", "good", and "ultimate" choices for shrinkage factors. We also develop an exact parallel between the fixed coefficient and the random coefficient formulations of minimum mean-squared-error shrinkage of a single parameter.

4.1 Classical "Optimal" Shrinkage
4.1.1 Diagonal Elements of Mean Squared Error Matrices
4.1.2 MSE Measures Depending Only Upon Diagonal Elements
4.1.3 Weighted Mean Squared Error Measures
4.1.4 The MSE in Specific Directions
4.1.5 Balancing Components of MSE Parallel to and Orthogonal to the Unknown True Coefficient Vector
4.1.6 Canonical Form for Optimal Shrinkage of a Single Fixed-Effect Coefficient
4.2 Classical "Good" Shrinkage
4.3 Classical "Ultimate" Shrinkage
4.4 Random Coefficient Shrinkage
4.4.1 A Within-Batch and Between-Batch Variation Model

4.4.2 Canonical Form for Optimal Shrinkage of a Single Random-Effect Coefficient
4.5 Summary

5. NORMAL-THEORY MAXIMUM LIKELIHOOD: BLUEs and BLUPs

This chapter includes a review of the basic theory of BLUEs and BLUPs. For shrinkage in fixed coefficient models, we develop not only general methods for identifying maximum likelihood estimators within arbitrary shrinkage families but also **closed form** expressions for optimal shrinkage estimators within the 2-parameter family. For shrinkage in random coefficient models, we discuss why maximum likelihood estimates are rarely "linear" or "unbiased". And we review the maximum likelihood methods of Golub, Heath, and Wahba(1979) and of Shumway(1982) for the special case of completely random models with a single variance component.

5.1 Unrestricted Maximum Likelihood and BLUE Theory
5.2 The Likelihood of Mean Squared Error Optimality
5.2.1 Unrestricted Maximum Likelihood Shrinkage: The Cubic Estimator
5.2.2 Maximum Likelihood UNIFORM Shrinkage
5.3 Closed Form Expressions within the 2-Parameter Family
5.3.1 The most-likely-to-be-mse-optimal shrinkage extent, k, for given shape/curvature.
5.3.2 The most-likely-to-be-mse-optimal shrinkage shape/curvature, Q.
5.3.3 The limit as the shrinkage shape/curvature, Q, approaches $-\infty$.
5.3.4 Large Sample Chi-Squared Tests of MSE-Optimality
5.4 Maximum Likelihood Methods for Mixed Linear Models
5.5 Completely Random Models with a Single Variance Component
5.5.1 Demonstration that BLUP estimates are shrinkage estimates in this case.
5.5.2 Random coefficient maximum likelihood choice of shrinkage extent.

6. RISK (MEAN SQUARED ERROR) ESTIMATION and SIMULATION

In this chapter, we display normal-theory maximum likelihood estimates of scaled (or relative) mean-squared-error (MSE) risk. In addition to estimates of risk in individual coefficients, we consider estimates for arbitrary linear combinations. And we explore corrections for bias and "range." Then we examine Monte-Carlo simulation results showing that reduced MSE risk is easier to actually achieve when coefficients are random than when they are fixed.

6.1 Stein's Unbiased Estimate of Overall Predictive Risk
6.1.1 Contraction Towards a Linear Variety
6.1.2 Minimum Mean Squared Error Estimation of σ^2
6.1.3 Stein Contraction Formulas
6.2 Estimates of Shrinkage Risk: Fixed Coefficient Cases
6.2.1 Unbiased Normal-Theory Estimates
6.2.2 Correct-Range Estimates

6.2.3 Shrinkage Factors Minimizing Scaled Risk Estimates
6.2.4 The Estimated Risk in Arbitrary Linear Combinations
6.2.5 Mallows-like Estimates of Predictive Mean-Squared-Error
6.3 Estimates of Shrinkage Risk: Random Coefficient Cases
6.4 Monte-Carlo Risk Simulation
6.4.1 Simulated Risk for Fixed Coefficient Models
6.4.2 Simulated Risk for Random Coefficient Models
6.4.3 Summary of Risk Simulation Results

7. RANDOM COEFFICIENT FORMULATIONS

Here we review iterative methods (Newton-Raphson, Fisher Scoring, EM, REML) for solving Henderson's "mixed model equations." And we detail specific applications in which random-coefficient models are more realistic (and provide "better" estimates) than fixed coefficient models.

7.1 Estimation of Random Effects
7.2 Estimation of Variance Components
7.3 Variation Between and Within Production Batches
7.4 Pharmaceutical Stability Models

8. BAYESIAN FORMULATIONS

We review the normal-theory, hierarchical models of Lindley and Smith(1972) as well as the empirical Bayes formulation of Efron and Morris(1977). We also discuss exactly why and how the Bayesian variance of a shrinkage estimate exceeds its classical variance. We derive both Theil's measure of the proportion of posterior precision due to sample information and also Shannon's measure of information gain. Finally, we comment on proposals for nonconjugate Bayes formulations.

8.1 Bayesian Conjugate-Normal Linear-Model Formulations
8.2 Bayesian Diagnostic Checking
8.3 More Bayes' Measures of the Extent of Shrinkage
8.4 Nonconjugate Bayes Formulations
8.5 An Empirical Bayes Likelihood Approach

9. COMPUTATIONALLY INTENSE METHODS

We start with two sections on “errors-in-variables" models under which least-squares estimates are biased and inconsistent; we first show how random data un-rounding can suggest shrinkage to improve stability of estimates, but we also explore maximum likelihood methods for multivariate normal “structural" models that suggest expansions. Next, we review iterative methods which, although traditionally applied to regressor variable subsetting or robust fitting, can also be used in shrinkage regression estimation. Specifically, we discuss methods of cross-validation for choice and assessment of shrinkage and methods that down-weight certain types of otherwise “influential" observations.

9.1 Data Perturbations After the Last Decimal Place and the Perturbation-Limit
9.2 Multivariate Normal Errors-in-Variables Models
9.3 Cross-Validation, Bootstrapping, and Sample Reuse Methods
9.4 Iterative Re-Weighting Methods

10. TOPICS of HISTORICAL INTEREST, HEURISTIC ARGUMENTS, and COMMON MISCONCEPTIONS

We start with a review and critique of the many contributions of Art Hoerl and Robert Kennard to the theory and practice of **ridge regression**. But a wide spectrum of alternative approaches are also described.

10.1 The Contributions of Hoerl and Kennard
10.2 The Obenchain-Vinod “chain-rule-argument"
10.3 Methods based upon Fictitious Data Augmentation
10.4 Preliminary Test Methods for imposing linear restrictions or detecting multicollinearity
10.5 Methods for Relaxing Correlations among Coefficients
10.6 Methods Utilizing Estimates of One.

Part Two: Shrinkage Regression Applications and Implementations

11. TRACE DISPLAYS: THE PSYCHOLOGY OF PERCEPTION

We start with a discussion of the many advantages of adopting the Multicollinearity Allowance (MCAL) choice for scaling along the horizontal (shrinkage extent) axis of a TRACE display.

Then we explain not only how to "read" the 5 major types of ridge TRACE displays but also how to focus on specific details using **Path Projection** onto two-dimensional linear subspaces.

> "The greatest thing a human soul ever does in this world is to SEE something, and tell what it SAW in a plain way. ... To see clearly is poetry, prophecy and religion - all in one."
>
> John Ruskin
> Modern Painters, 1888

11.1 Multicollinearity Allowance (MCAL) Scaling
11.1.1 Alternative Measures of Shrinkage Extent
$MCAL = k \cdot trace[\,(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} + k \cdot \boldsymbol{I}\,)^{-1}]$
11.1.2 Generality and Comparability
11.1.3 Finite Range
11.1.4 Stable Relative Magnitudes Plot as Straight Lines
11.1.5 Bayesian Posterior Precision Interpretation
11.1.6 Rank Deficiency Interpretation
11.2 TRACE Displays of Shrinkage Coefficients
11.3 TRACE Displays of Shrinkage (Delta) Factors
11.4 TRACE Displays of Estimated MSE in Individual Coefficients
11.5 TRACE Displays of Excess MSE (OLS minus Ridge) Eigenvalues
11.6 TRACE Displays of the Inferior Direction Associated with Excessive Shrinkage
11.7 Path Projection Onto Two-Dimensional Linear Subspaces
11.7.1 Fitted Coefficient and Likelihood Hyperellipsoid Projections
11.7.2 Inferior Direction Animation
11.7.3 Regressor Coordinate Projections
11.8 Advantages of One- and Two-Parameter Families Over Unrestricted Shrinkage
11.8.1 Invariance Arguments for Shape QPAR $= +1$
11.8.2 Minimax Arguments for Shape QPAR $= +1$ or $+2$
11.8.3 MSE Reduction Potential Arguments for Shape QPAR $< +1$
11.8.4 Geometric Arguments for Shape QPAR ≤ 0
11.9 Shrinkage Estimates should NOT be Significantly Different from Least-Squares

12. USAGE of RXridge

RXridge performs normal-theory maximum likelihood computations for 2-parameter generalized shrinkage regression. It features user-friendly windows, menus, graphics, and visual review of ridge files that have been written to disk. RXridge handles **exact** (numerical) singularities as well as nearly **multicollinear** (statistically ill-conditioned) examples.

13. USAGE of RXtraces

RXtraces creates **interactive** CGA graphics displays of all 5 types of ridge TRACES ...including estimates of (i) regression coefficients, (ii) scaled mean squared errors, (iii) excess (ordinary-least-squares minus ridge) eigenvalues, (iv) inferior direction cosines, and (v) the shrinkage factor pattern. RXtraces is ideal for ridge **training** applications (workshops and/or self-paced learning) in which students simply review RXridge or RelaxR outputs across a spectrum of previously-computed numerical examples.

14. USAGE of PathProj

PathProj creates **interactive** CGA graphics displays of the projection of ridge shrinkage path statistics (regression coefficients, inferior direction, regressor coordinates, etc.) onto any 2-dimensional linear subspace (specified via orthogonal direction cosine vectors.) Literally **WATCH** as an inferior direction first “appears" and then changes its orientation as you step along a ridge shrinkage path.

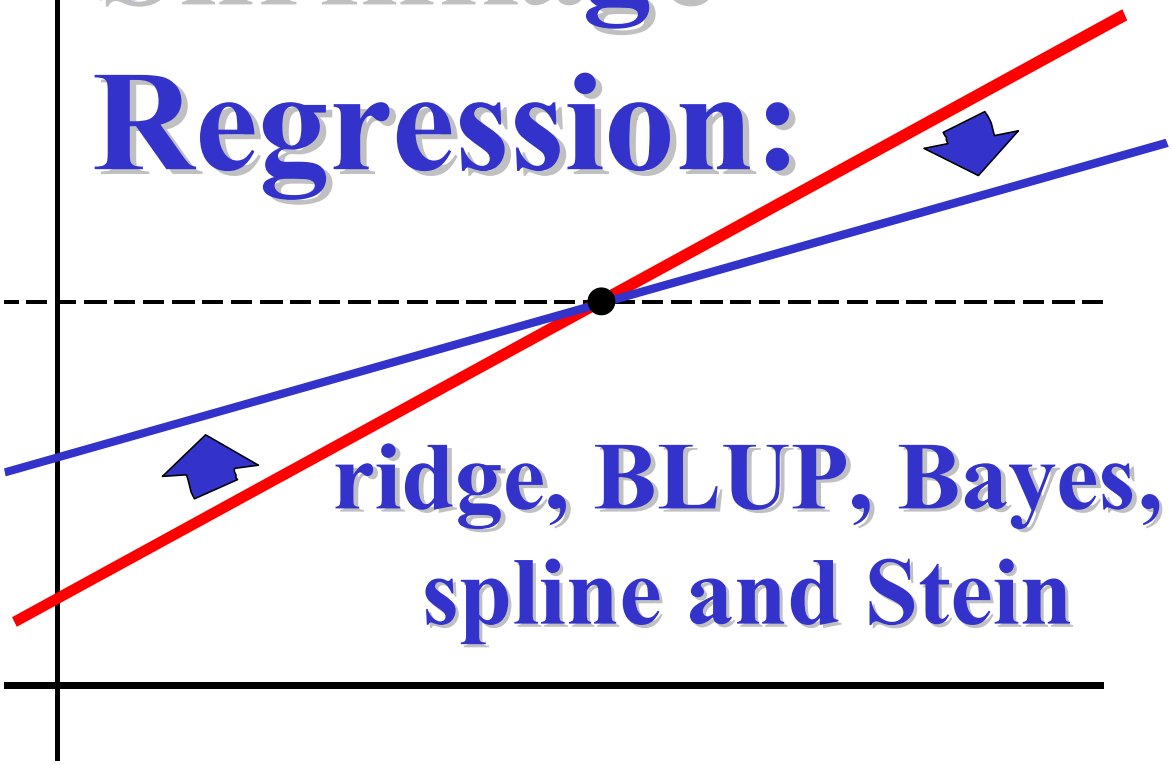
15. USAGE of RXrisk, RXshrink, and RXmsesim

16. CASE STUDY 1: PORTLAND CEMENT (MIXTURE) DATA

17. CASE STUDY 2: GASOLINE MILEAGE DATA

18. CASE STUDY 3: PHARMACEUTICAL SHELF LIFE ESTIMATION

Shrinkage Regression:



Chapter 01: Introduction

Bob Obenchain, Ph.D.
softR$_X$ freeware
13212 Griffin Run
Carmel, Indiana 46033-8835

Chapter 1: INTRODUCTION

Modern regression methodology and terminology trace their history back at least as far as the work of Carl Friedrich Gauss and Adrien Marie Legendre at the start of the nineteenth century on fitting orbits to astronomical data. The first published description of the "principle of least squares" fitting was that of Legendre(1805) who coined its name (moindre carrés.) But Gauss apparently used this method routinely for "combination of observations," starting as early as 1795. And it was Gauss(1809) who made the initial contributions to the least squares **theory of estimation**, including the normal distribution of "errors" and the most basic foundations for maximum likelihood.

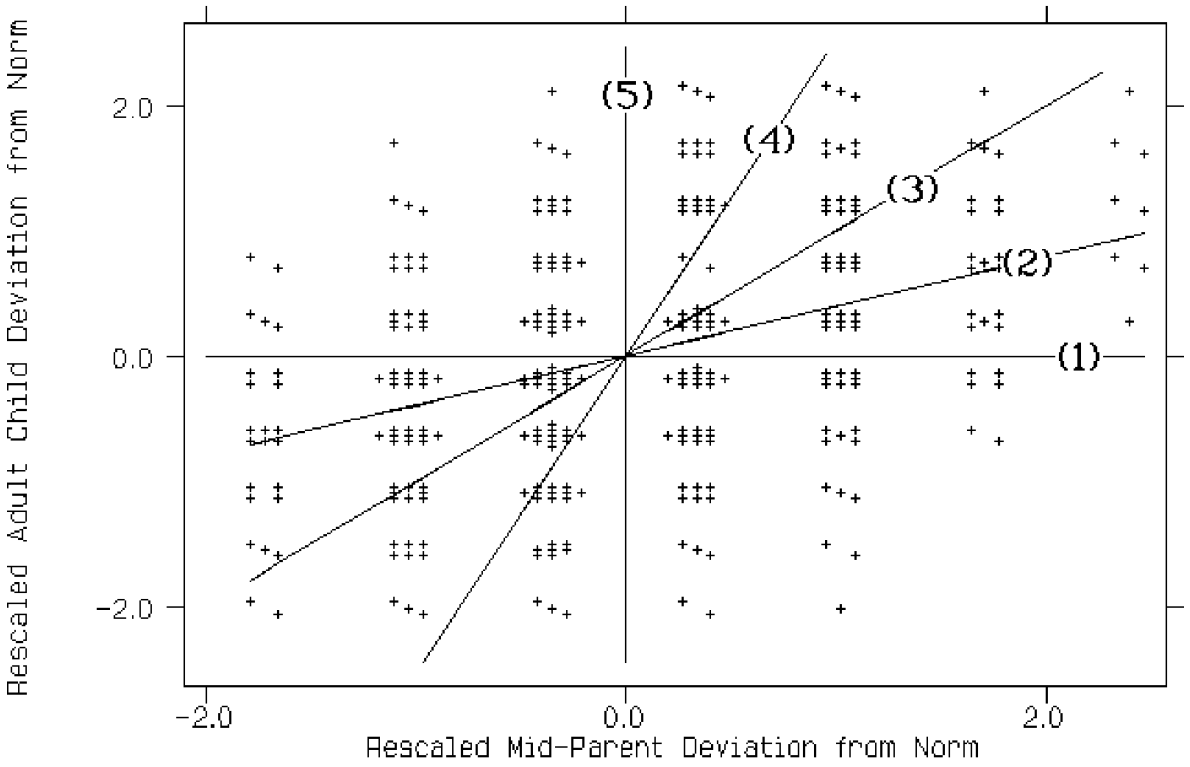
1.1 Galton's "Shrinkage" Interpretation of Regression

The next major contributor to least squares theory/terminology was Francis Galton, who published several papers related to least squares fitting in the late 1800's. Galton(1877) proposed a numerical measure, originally called "reversion," to quantify relationships between physical characteristics of parents and their offspring. Galton's measure expressed the expected value of a child's deviation from the norm as a fraction of its parent's deviation. Galton used data on the size of sweet pea seeds from mother and daughter plants in his 1877 paper, apparently because he had no suitable data from human populations at that time. Eight years later, Galton(1885) tabulated the height of 329 adult children versus the mid-height of their parents, giving counts in inch wide cells. It was also in this 1885 paper that Galton (with help from Cambridge mathematician Hamilton Dixon) not only laid the foundation for the modern concepts of bivariate normal constant-density ellipses and the correlation coefficient but also hinted at principal components. But it was, perhaps, Galton's "regression" terminology in this 1885 paper, "Regression Towards Mediocrity in Hereditary Stature," that has made the greatest long-term imprint on our subject.

The scatter-plot of Figure 1.1 illustrates Galton's motivation for describing least squares fits in terms of a "regression" or "shrinkage" towards mediocrity, where mediocrity is represented here by the sample mean height. Figure 1.1 displays the data from Galton(1885), Table I, using patterns of points to represent counts in the 58 cells of Galton's classification; the vertical (Y) coordinate of each point represents an adult child height, while the corresponding horizontal (X) coordinate is the mid-height of that child's parents. We have placed the origin of coordinates at the data centroid, we have scaled the X and Y axes in units of the corresponding sample standard deviations, and we have used the Pearson-product-moment formula to compute the

sample correlation (0.397) between parent and child heights. (Galton centered his bivariate ellipses at the median height, and scaled axes using observed ranges.)

Figure 1.1 Galton's Regression Towards Mediocrity



The five lines that pass through the data centroid, (0,0), in Figure 1.1 have the following properties:

(1) The horizontal line represents the extreme case for prediction of Y from X in which the predicted adult-child height is the mean height for all values of parent mid-height.

(2) The line with slope 0.397 represents the least squares regression of adult-child heights onto the mid-height of their parents. This is the best fitting line in the sense that the sum-of-squares of **vertical** deviations from the fitted line is minimized.

(3) The 45° line (slope of 1) represents the first principal component axis of the data. This is the best fitting line in the sense that the sum-of-squares of deviations measured **orthogonal** to the fitted line is minimized.

(4) The line with slope 1/0.397 = 2.52 represents the least squares regression of parent mid-height onto adult child height. This is the best fitting line in the sense that the sum-of-squares of **horizontal** deviations from the fitted line is minimized.

(5) The vertical line represents the extreme case for prediction of X from Y in which the predicted parent mid-height is the mean height for all values of adult child height.

If one had to use a single line to predict not only Y from X but also X from Y, it would seem to make the most sense to use the first principal component axis line, numbered (3) in Figure 1.1. When we say that the same line would be used for both predictions, we mean specifically the following: if $y = \alpha + \beta \cdot x$ is used to predict y from x, then the equation for predicting x from y is of the form $x = (y - \alpha)/\beta$ whenever $\beta \neq 0$.

Rather than restrict attention to a single line, suppose instead that one can use a different line to predict Y from X than that used to predict X from Y. Under this supposition, Galton observed that the best prediction of Y from X is line number (2) in Figure 1.1, which represents a certain "shrinkage" (or rotation) of line (3) towards line (1). Similarly, the best prediction of X from Y is line number (4) in Figure 1.1, which represents a corresponding "shrinkage" of line (3) - but this time towards line (5). Galton's observation of this "shrinkage" apparently was his motivation for referring to the least squares fits as "regression" lines.

The modern shrinkage regression methods discussed in this book usually suggest **more** shrinkage than least-squares. These methods yield a fitted Y-on-X line between lines (1) and (2) and a fitted X-on-Y line between lines (4) and (5). We will explore a wide variety of arguments suggesting that this sort of "extra" shrinkage is well worth our careful consideration when fitting regression models to ill-conditioned data. On the other hand, we also explore "errors-in-predictor-variables" arguments in section §9.2 that suggest **less** shrinkage than least-squares!

1.2 The Primary Theme of This Book

Development of shrinkage regression methodology was my primary statistical research interest for about 19 years (1974-1992.) This manuscript surveys just about all of the inference tools that I, personally, have found to be both theoretically sound and practically useful in multiple regression modeling. A striking thing about these diverse techniques is that they all seem to reinforce a common theme, each from its own, unique point-of-view. And that common theme is:

> Estimating the regression coefficients of a multiple regression model when the available regressor data are ill-conditioned (intercorrelated) is nothing less than **a full-blown problem in multivariate analysis**.

This wasn't the point-of-view I started out with in 1974 when I began examining shrinkage methodology applicable to regression models. Back then I was already familiar with the elegant result of James and Stein(1961), the observation of Lindley(1962), and the work of Sclove(1968). And I was reading the extremely optimistic claims of Hoerl and Kennard(1970a,b), and other "pragmatic" regression practitioners seemed to be agreeing with them [McDonald and Schwing(1973), Marquardt and Snee(1975), etc., etc.] The fundamental "shrinkage" message back then seemed to be that least-squares regression coefficient estimates

really weren't very good, especially when the given regressor data were ill-conditioned (nearly multi-collinear), and that there were some relatively “obvious" ways that least-squares could be improved upon.

From today's perspective, the shrinkage “revolution" of the 1970's was a rather dismal failure. Least-squares estimation has retained its rightful place of prominence in the modern repertoire of multiple regression methodologies. On the other hand, modern personal computer software systems probably are encouraging statistically “naive" users to reach new heights in the “abuse" of least-squares; see comments in Box(1970), Dempster(1973) and Tukey(1975). [We statisticians really could help software developers out by reaching some sort of consensus about which methods are not only moderately efficient but also robust/resistant enough for almost “reckless" use by non-statisticians.]

On the other side of the same coin, modern personal computer software systems also enable/encourage true data visualization and revelation of anomalies “hidden" within data. The “hot" methodologies of today are, perhaps, transformation of variables, predictive sample reuse (jackknife and bootstrap), regression diagnostics (outliers, leverage & influence) and robust/resistant methods. Multiple regression “veterans" have learned the value of graphical displays in examining their data and models, and now they have some really good hardware/software **tools** to do exactly that!

There are sound, theoretical reasons why shrinkage methodology failed to deliver any sort of “knockout punch" to least-squares in the 1970's:

> Bunke(1975) and Brown(1975) established that least-squares estimates are minimax (admissible) when the risk function is multivariate (matrix valued.) Guaranteed ways to realistically “beat-the-system" on long range average **cannot exist** without “added" information that usually isn't available! And Obenchain(1977) argued that classical, normal-theory hypothesis tests and confidence regions based upon shrinkage estimators are actually identical to “unbiased" least-squares tests and regions of the same statistical confidence.
>
> Systematic examination of fitted residuals is an essential phase of regression modeling. Obenchain(1975a) showed that least-squares residuals have optimality properties that assure their usefulness in these sorts of exploratory tasks **even when they are based upon incorrect expectation and/or dispersion models**. And today's most efficient, robust/resistant multiple regression methodologies are apparently the ones based upon iterative re-weighting of least-squares residuals, Andrews(1974).

Why, then, am I writing a book about shrinkage regression now that the shrinkage “revolution" has failed? My answer is that statisticians need to be well informed about how the **evolution** of shrinkage regression is still continuing today!

The future for applications of shrinkage methodology to multiple regression models lies primarily in their unique ability to highlight and clarify results from least-squares analyses. If shrinkage practitioners can learn to consistently apply their methods with subtlety and

understanding, respect within the statistical community for shrinkage approaches will grow naturally with time. For example, certain effects which are quite large **numerically** may not be significantly different from zero **statistically** when the available regressor data are ill-conditioned. Methods that focus shrinkage upon exactly these kinds of "noise artifacts" help us avoid potential misinterpretations of data.

Besides, I believe that fundamentally sound ideas tend to be characterized by the "intellectual robustness" property that they can be motivated from many different but mutually reinforcing points-of-view. And I am writing here about multiple facets of shrinkage regression methodology precisely because I feel that these approaches are sound in that sense.

As an example of a futuristic use for shrinkage regression, consider the following scenarios...

IF THERE WERE ONLY ONE PREDICTOR VARIABLE

The task of regressing a single response variable (Y) onto a single predictor variable (X) is not particularly perplexing, especially with modern computer hardware and software. But this is the case only because we know exactly which plot to make in order to **SEE** literally everything that is happening! Specifically, we plot the available data as points on the X - Y plane, and we then simply superimpose candidate regression lines (or curves.) For each candidate fit, we immediately **SEE** outlying response values, high leverage regressor points, lack-of-fit, patterns in residuals, variance-bias-tradeoffs, etc., etc. Yes, we may wish to augment this X - Y plot with other plots, say, probability plots of residuals to check distributional assumptions. But our bottom line on this "one response, one predictor" special case is simply this:

We know we can proceed with great confidence!

Next...

IF THERE WERE ONLY TWO PREDICTOR VARIABLES

If we have access to sufficiently powerful computer hardware/software for 3-dimensional displays, we can retain much of our comprehensive visual insights when a single response is fit by two predictor variables.

WHEN THERE ARE "MANY" PREDICTOR VARIABLES

Our abilities to routinely visualize fits with three or more Xs and one Y (let alone several Ys!) are meager at best. What we **SEE** in these cases depends almost totally on which linear combinations we (unilaterally) decide to use as axes for plots. And our chances of "accidentally" making the "right" plots (and thereby gain insights about how several regressor variables might interact in predicting the response variable) decrease rapidly as the number of potential regressor variables increases. Estimates of regression coefficients in multiple regression models are intercorrelated precisely because the available regressors are intercorrelated; marginal distributions and bivariate plots simply do not reveal **all** that we need to know about complicated multivariate interdependencies.

IF WE FORM A COMPOSITE PREDICTOR VARIABLE

Of course, if we have a tentative estimate, $\mathbf{b}^{\bigstar}$, for a vector of regression coefficients, β, then we can form a new, composite regressor variable. Specifically, consider $\mathbf{X}^{\bigstar} = \mathbf{X}\,\beta^{\bigstar}$, which will be a single column vector when $\beta^{\bigstar}$ is, say, a vector of **unit length** parallel to $\mathbf{b}^{\bigstar}$. This, in turn, allows us to consider the **simple regression** of Y onto this **single** $X^{\bigstar}$ variable. This allows us to return to an efficient and familiar mechanism to **SEE** what is happening: Are there outlying response values? Which regressor combinations have highest leverage on this fit? Is there obvious lack-of-fit? Are there patterns in residuals? And, are there interesting possibilities for variance-bias-tradeoffs? All of these sorts of insights become suddenly available once we form a composite $X^{\bigstar}$ regressor variable, draw the implied bivariate Y versus $X^{\bigstar}$ plot, and superimpose fitted line(s). This tactic of visualizing the simple regression of our response variable onto a tentative, composite regressor will be termed **Visual Re-Regression** or VRR.

We should not be so naive as to think that VRR is any sort of panacea or even a totally reliable heuristic that can never mislead us. Rather than actually untangle a "Gordian Knot" of multivariate complexity, we are simply considering one of many possible ways to "cut" through it. Furthermore, it would seem (to me) to be quite time consuming, tedious and wasteful of human and/or computer resources to ever attempt to consider **all** possible orientations for the $\mathbf{b}^{\bigstar}$ vector, yielding a continuum of VRR scenarios. If VRR tactics are to become an important part of a comprehensive, overall regression strategy, practitioners will need to have fairly straightforward ways to generate only a **few, interesting** tentative estimates, $\mathbf{b}^{\bigstar}$, for detailed VRR study.

CONTENTION: The shrinkage regression techniques considered in this book are sound mechanisms for generating tentative $\mathbf{b}^{\bigstar}$ estimates worthy of detailed VRR study. New visualization tools, such as TRACES and projection plots, empower us to implement shrinkage regression methodologies, thereby revealing distortions in the relative magnitudes of fitted coefficients that are due to intercorrelation among regressors. Once we have exploited potential variance-bias-tradeoffs to untangle multicollinearities and form tentative $\mathbf{b}^{\bigstar}$ estimates, we can

then return to the more familiar paradigm of VRR for further exploratory and confirmatory analyses.

1.3 How are Shrinkage Regression Methods Typically Applied?

Let us consider a thumbnail sketch of how shrinkage regression techniques might be applied to a simple numerical example. In the spirit of a PREVIEW-OF-COMING-ATTRACTIONS, we will now freely use concepts and terminology that have not, as yet, been fully motivated and explained! Our objective is simply to give the reader of this introduction a little bit of the ultimate "touch and feel" associated with confidently applying modern shrinkage regression techniques to an ill-conditioned dataset.

We will analyze data from 5 of the 11 variables on all 32 automobiles used by Hocking(1976) to illustrate analysis and variable selection techniques in regression. The response variable of interest will be MPG = miles per gallon, and our four predictor variables will be:

CYLNDS = number of cylinders

CUBINS = cubic inches of engine displacement

HPOWER = engine horsepower

WEIGHT = total auto weight in pounds

We chose this numerical example to illustrate shrinkage regression methodology in the hope that many readers may be ready-and-willing to harbor pre formed opinions about the effects of these four factors on gasoline mileage. In fact, I hope that you will agree with me that gasoline mileage "should" **decrease** as any of our four basic factors is **increased** ...RIGHT?

Hocking(1976) used readily available data from three 1974 issues of **Motor Trends** magazine, not results from any sort of "designed experiment" or "representative cross-section" of automobiles. As a result, his regressor data are ill-conditioned (highly intercorrelated) and don't do a very good job of reaching into the "corners" of 4-dimensional predictor space. In fact, Henderson and Velleman(1981) point not only to curiosities in the coding of the CYLNDS variable but also to potential biases due to inclusion of data on 7 Mercedes (including one diesel) and 6 mid-engine/sports cars. Furthermore, they show that examination of preliminary plots of MPG versus the potential regressors reveals "curvatures" which strongly suggests that GPHM (Gallons Per Hundred Miles) would be much more nearly linearly related to the given regressors than is MPG. In fact, they also suggest using "an understanding of cars in this collection" to create a new variable, HPOWER/WEIGHT, as a "measure of how overpowered a car is."

Although the general sorts of preliminary graphical strategies/tactics used by Henderson and Velleman(1981) on this dataset are **highly recommended** by this author, they do not necessarily

represent the very **first** things that one might try out. Yes, cautious practitioners will always want to “look" at their data before “leaping" to model fitting. But it seems to me, at least, that nothing short of an actual attempt at fitting a model to his/her data can give a pragmatist cause to doubt that even the most simple and straightforward approach will “work." Therefore, let us pretend here that we are naively unaware that we are working with a subset of an “infamous" dataset. And let us use shrinkage regression methodology in our “first," exploratory attempt at fitting a model to the gasoline mileage data.

Examination of summary statistics that describe the “coverage" (or “spread") of the available predictor variable combinations is typically the starting point in practical applications of shrinkage regression. We first center regressor coordinates at their mean values and rescale each variable in units of its own standard deviation. The table below (Table 1.1) presents computational results for the principal component rotation of regressor coordinates. We see (row four) that the largest single numerical contribution (-0.6094) to the least squares regression coefficient vector comes from the dimension of regressor space that is least adequately “covered" by the available data; the regressor standard deviation along that last (minor) principal axis is only 1.429 standard units compared with 10.31 along the first (major) principal axis. In other words, the available data are extremely ill-conditioned.

Table 1.1: Component Summary Statistics

Singular Values	Uncorrelated Components	Principal Correlations	t-Statistics
10.31	-0.4872	-0.9025	-12.05
3.35	-0.1325	-0.0797	-1.06
2.084	0.1535	0.0575	0.77
1.429	-0.6094	-0.1564	-2.09

Because of this ill-conditioning, we should NOT be surprised when undesirable features emerge in the following table of least-squares statistics.

Table 1.2: Ordinary Least Squares Summary

Marginal Correlations		Regression Coefficients	Relative Std.Errors	t-Statistics
-0.8522		-0.3832	0.4662	-1.97
-0.8476	<-->	0.2385	0.5785	0.99
-0.7762		-0.2336	0.3315	-1.69
-0.8677		-0.6257	0.3955	-3.80

A “sign-conflict" has arisen between the marginal correlation of the second regressor (CUBINS) with the MPG response and the corresponding regression coefficient. Naturally, we wonder if we can resolve this conflict via shrinkage regression methodology. But which

“shape" of shrinkage should we use? (Uniform shrinkage is of no real interest here because coefficient two would then remain positive while the other three would all remain negative as all four are shrunken to zero.) When in doubt, why not let the data themselves suggest an appropriate “shape" and “extent" of shrinkage? The following table summarizes results using closed form expressions for the coefficient vector most likely to be mean-squared-error-optimal (under normal distribution theory) along a spectrum of nine possible shrinkage-shape paths ...

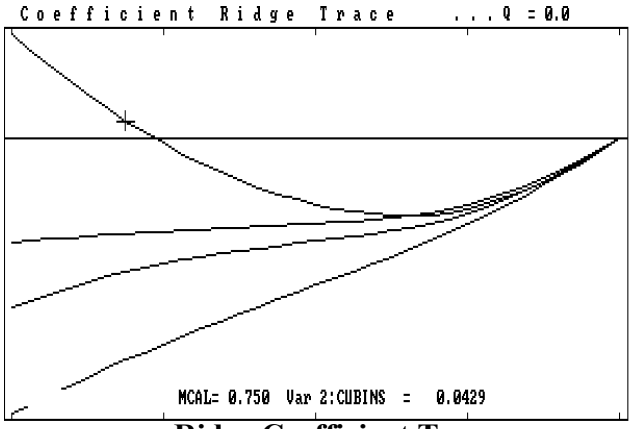
Table 1.3: Shrinkage Shape Summary

QPAR	MCAL	Konst	CRLQ	Chi-Sq	Best
2.00	2.717	0.314	0.2980	57.91	
1.50	1.839	0.267	0.4284	55.00	
1.00	0.733	0.224	0.6492	46.25	
0.50	0.413	0.303	0.8681	27.77	
0.00	0.611	1.01	0.9670	9.94	
-0.50	1.340	7.42	0.9881	4.00	
-1.00	2.121	74.1	0.9881	3.97	<<<
-1.50	2.601	783	0.9853	4.86	
-2.00	2.850	8.27e+3	0.9830	5.54	

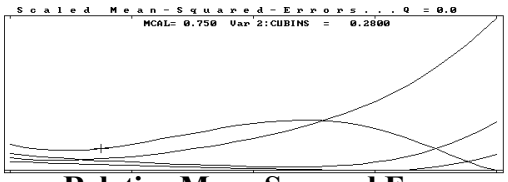
This analysis favors a shrinkage extent of about MCAL $= 2$ along the path of shape QPAR $= -1$. And we have already ruled out the uniform (QPAR $= +1$) shape as being of no interest in this application. Suppose we decide to actually explore the path of shape QPAR $= 0$ because (i) this corresponds to the well known special case of “ordinary" ridge regression, Hoerl and Kennard(1970a,b), and (ii) less shrinkage may be needed (MCAL $= 0.6$) along this path than along the QPAR $= -1$ path.

Next, we generate and examine the 5 major ridge TRACE displays for the QPAR $= 0$ path...

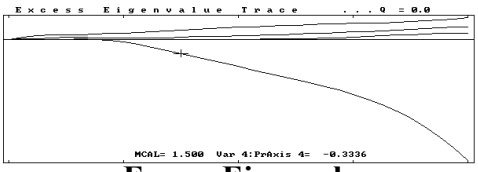
Figure 1.2: Shrinkage TRACE Displays



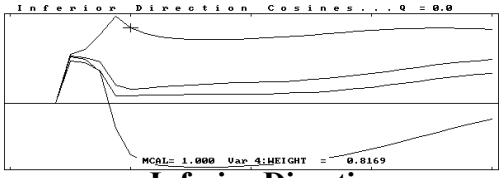
Ridge Coefficient Trace



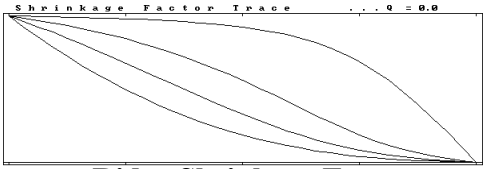
Relative Mean Squared Errors



Excess Eigenvalues



Inferior Direction

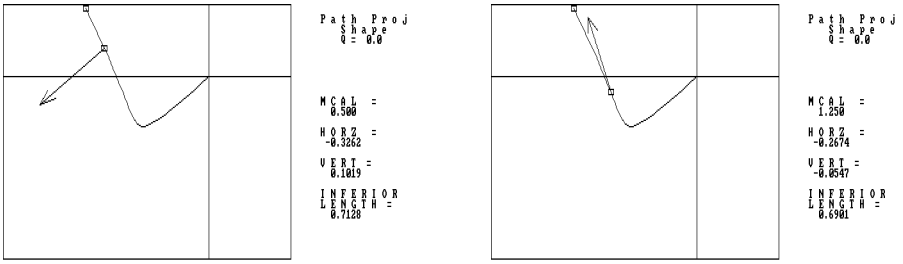


Ridge Shrinkage Factors

Many details of TRACE displays in Figure 1.2 are really too small for you to see very clearly. And we haven't yet discussed how to read them! So let us simply observe the following: All four ridge coefficients can be made negative by shrinking to at least MCAL = 1.0 along the QPAR = 0 path. However, that appears to be considerably more shrinkage than can be justified in terms of reduction in classical (fixed coefficient) mean-squared-error.

Additional insight is provided by combining information from the coefficient and inferior-direction traces via **Path Projection** onto the plane spanned by regressors 1 and 2 (CYLNDS and CUBINS). What we observe in Figure 1.3 (below) is that, just as the CUBINS coefficient switches from positive to negative in sign, the inferior direction abruptly changes orientation. In fact, it suddenly starts pointing **backwards** towards the least squares solution. This is a clear signal that shrinkage sufficient to make the CUBINS coefficient negative is actually **excessive**!

Figure 1.3: Shrinkage Path Projections



Maximum likelihood calculations following the empirical Bayes approach of Efron and Morris(1977) or the random coefficient approach of Golub, Heath, and Wahba(1979) and Shumway(1982) are also possible, of course. On the other hand, because there are no closed form solutions for the estimators that minimize the corresponding negative-log-likelihoods, these statistics must be calculated at a mesh of "steps" along the shrinkage path of the desired shape (QPAR = 0, here):

Table 1.4: Shrinkage Extent Monitoring

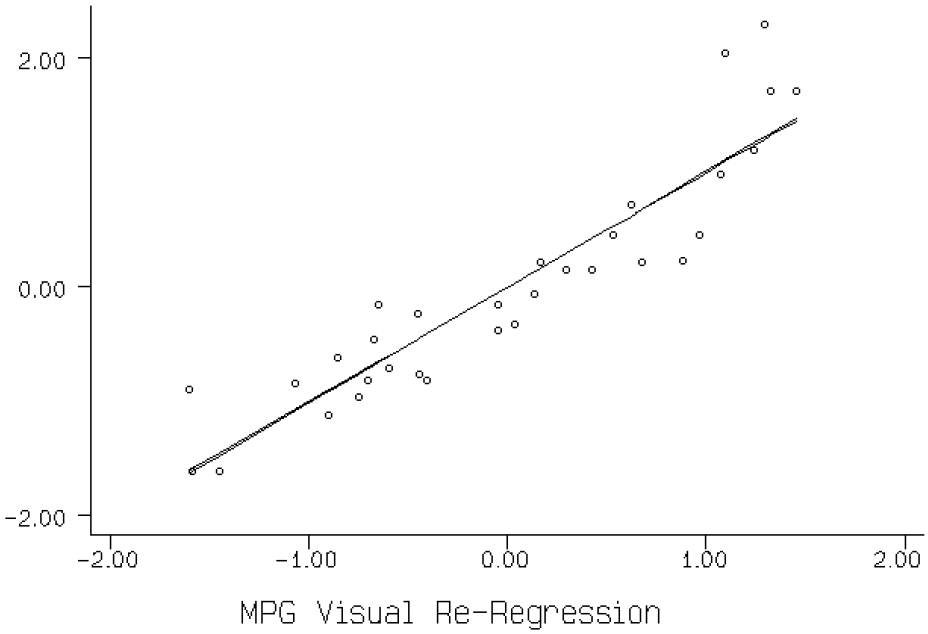
MCAL	CLIK	EBAYS	RCOEF
0.000	+infinity	+infinity	+infinity
0.125	90.9	17.3	17.4
0.250	29.6 <(2/R)	15	15.2
0.375	15.1	13.9	14.1
0.500	10.7	13.3	13.6
0.625	9.95 <<<	12.9	13.3
0.750	10.7	12.8 <<<	13.2 <<<
0.875	12.1	12.8	13.2
1.000	13.9	13	13.4
2.000	30.6	19.6	18.6
3.000	45.8	52.5	35.4
4.000	60.4	151	60.4

Note that both the empirical Bayes and the random coefficient maximum likelihood criteria also favor less shrinkage than is necessary to produce a negative CUBINS coefficient. And the (2/R)ths Rule-of-Thumb [where R=4 here] would limit shrinkage to only MCAL $\leq$ 0.3. As a compromise, let us proceed using the QPAR=0 and MCAL=0.625 solution, (– 0.3143, +0.0716, – 0.2190, – 0.5235), instead of the least-squares (MCAL=0) solution, (– 0.3832, +0.2385, – 0.2336, – 0.6257).

The final "sanity-check" phase of our analyses would then consist primarily of Visual Re-Regressions (VRR) using the composite regressors defined using (– 0.3143, +0.0716, – 0.2190, – 0.5235) and/or resulting from elimination of CUBINS from our model, namely

(– 0.279022, 0.0, – 0.205202, – 0.514149). On the other hand, VRR may reveal serious problems with outlying response values, high leverage regressor points, lack-of-fit, etc. that would cause us to "set aside" some of our available observations and/or transform variables. In this case, we essentially have to start over, almost from scratch!

Figure 1.4: Visual Re-Regression



To keep our exposition of typical shrinkage regression applications rather brief, we will make only a few simple observations at this time:

Our VRR (Figure 1.4) for the [QPAR=0, MCAL=0.625] shrinkage solution $\mathbf{b}^{\bigstar} = (-0.3143, +0.0716, -0.2190, -0.5235)^{\mathbf{T}}$ demonstrates close agreement (in both predictions and residuals) between shrinkage-regression and the ordinary-least-squares fit to the gasoline mileage data. Major variance-bias tradeoffs have eluded us!

The parameterization we are working with is, in a quite pragmatic sense, not a very satisfactory one. Sufficient shrinkage to make the CUBINS coefficient negative apparently cannot be justified!

Our VRR displays the (possibly systematic) "curvature" noticed by Henderson and Velleman(1981) which prompted them to try an inverse transformation of the response variable, GPHM = 100/MPG.

In short, our exploratory analyses (including our attempt at VRR confirmation) should have convinced us that we have not yet developed any sort of totally satisfactory model. Unless we

have already exhausted all available time/resources, we should roll up our sleeves and start over, almost from scratch! [P.S. The GPHM response transformation really helps!!!]

1.4 Which Part of the Book Should I Read Next?

The remainder of the book is divided into two very different parts. Part One (Chapters §2 through §10) contains some rather heave reading on the general **Theory and Methodology** of Shrinkage Regression. Part Two (Chapters §11 through §18) contains some relatively light reading on **Applications and (Computer) Implementations** of Shrinkage Regression.

1.4.1 Do We Really Need All This NOTATION for CANONICAL ROTATION?

If you do start reading at the start of **Part One** of the book next, you will find that its first two chapters introduce a great deal of rather technical details...

Chapter §2: terminology/notation for linear statistical models and ill-conditioning.

Chapter §3: generalized shrinkage regression terminology/notation.

You may well ask "Why does our in-depth survey of shrinkage regression methodology have to start **so slowly** with a pair of chapters of technical details?"

After all, both Chapter §2 and Chapter §3 are rather long, and each introduces and describes a great deal of notation.

Therefore, let us stress exactly what we gain by adopting notation based squarely upon the principal axis rotation of regressor coordinates. Basically, there are three main motivations for our fundamental strategy:

(i) Rotation to CANONICAL FORM breaks regression coefficients down into their most elementary component parts, those of equation { 2.16 }. This establishes a parallel between the special case of uncorrelated regressors, { 2.7 }, and the general (intercorrelated regressors) case. It not only "unmasks" WRONG SIGNS problems but also eliminates possible confusion about distinctions between the NUMERICAL SIZE and the STATISTICAL SIGNIFICANCE of fitted regression coefficients. ILL-CONDITIONED multiple regression models are simply those that have relatively inadequate spread in given regressor coordinates along minor principal axes.

(ii) Principal axis rotation sweeps all variability onto the DIAGONAL of the variance-covariance matrix of the least squares (unbiased) regression coefficient estimates. When estimates are then defined in terms of shrinkage along these same principal axes, the off-diagonal elements of the resulting mean squared error matrix contribute to a rank=1 squares-and-cross-products structure for BIAS. This provides sufficient mathematical tractability to not only define an "optimal" shrinkage target along each axis but also to derive closed form

solutions for their normal-distribution-theory (restricted or unrestricted) maximum likelihood estimates.

(iii) Principal axis rotation greatly simplifies numerical computations. Generalized inverses of DIAGONAL matrices can be computed extremely quickly and accurately! Again, the canonical mean-squared-error matrices for generalized shrinkage estimators are always of a special form - namely, a diagonal matrix minus a symmetric, rank one matrix. Closed form expressions for the eigenvalues and eigenvectors of this type of matrix lead to extremely fast and accurate computations.

As we explore equation after equation in Chapters §2 and §3, we will actually be developing a uniform notational foundation for all of the basic concepts we will cover in our discussions. Later chapters will refocus our attention on these same fundamental relationships, giving alternative and enhanced motivations and interpretations.

1.4.2 What Topics are Covered in the Remainder of PART ONE?

In Chapter §4, we consider a wide variety of specific risk/loss functions that differentiate between desirable and undesirable forms/extents of shrinkage in regression. Most of these loss formulations lead to measures of mean-squared-error risk. These approaches invariably end up expressing desirable forms/extents of shrinkage as “target values," which are unknown because they are functions of unknown, “true" regression parameters.

In Chapter §5, we see how the classical, normal-distribution-theory likelihood function can be used to identify shrunken estimators most likely to be “good" or “optimal" in the senses of Chapter §4.

We consider maximum-likelihood estimates of risk in Chapter §6, along with modifications that make estimates unbiased or assure that they have “correct range."

Next we consider two important alternative formulations/motivations for shrinkage regression; we summarize random coefficient methods in Chapter §7 and empirical Bayes models in Chapter §8.

Chapter §9 discusses “computationally intensive" methods for resampling the available data and/or iterating towards an optimal solution that cannot be written as a closed-form expression.

Chapter §10 consists of nine sections on “miscellaneous" topics that are not really central to the arguments given in other chapters. Still, these topics are of historical interest, provide additional “heuristic" insights, or clarify common misconceptions about shrinkage regression methodology.

1.4.2 What Topics are Covered in PART TWO?

The materials presented in Part 2 of this book tend to focus much less on the "details" of mathematical theory and statistical methodology for shrinkage regression. Instead, we primarily focus our discussions in Part Two upon specific practical applications and on effective usage of personal computer software implementations for shrinkage regression.

Chapter §11 sets the exploratory, data analytic tone of Part 2 by exploring arguments about the PSYCHOLOGY-OF-GRAPHICAL-PERCEPTION that help one choose a scaling for the horizontal (shrinkage extent) axis on generalized ridge TRACE displays. All of our arguments here, from a spectrum of diverse points-of-view, seem to come down squarely on the side of using the "multicollinearity allowance," MCAL=R $- \delta_1 -$ ... $- \delta_R$, scaling along the horizontal axis of our ridge TRACE displays.

Chapters §12, §13, §14, and §15 describe usage of my **softRX freeware** (tm) systems for IBM-compatible (MS-DOS) Personal Computers to perform generalized shrinkage-regression calculations.

Chapters §16, §17, and §18 describe three CASE-STUDY numerical examples that illustrate how basic shrinkage strategy/tactics can be "dovetailed" in practical regression applications.

References for Chapter One

Andrews, D. F. (1974). "A robust method for multiple linear regression." **Technometrics** 16, 523-531.

Box, G. E. P. (1966). "Use and abuse of regression." **Technometrics**, 8, 625-629.

Brown, L. (1975). "Estimation with incompletely specified loss functions (the case of several location parameters.)" **Journal American Statistical Association** 70, 417-427.

Bunke, 0. (1975a). "Least squares estimators as robust and minimax estimators." **Math. Operations forsch u. Statist**. 6, 687-688.

Bunke, 0. (1975b). "Improved inference in linear models with additional information." **Math. Operations forsch u. Statist**. 6, 817-829.

Dempster, A. P. (1973). "Alternatives to least squares in multiple regression." **Multivariate Statistical Inference**. Eds. Kabe, D. G. and Gupta, R. P. Amsterdam: North-Holland Publishing Company, pp25-40.

Galton, F. (1877). "Typical laws of heredity in man." **Proceedings Royal Institute Great Britain**, 8, 282-301.

Galton, F. (1885). "Regression towards mediocrity in hereditary stature." **Journal Anthropological Institute**, 15, 246-263.

Gauss, C. F. (1809). "Theoria motus corporum coelestium." **Werke**, 7. (English translation: C. H. Davis, Dover, New York, 1963.)

Henderson, H. V. and Velleman, P. (1981). "Building multiple regression models interactively." **Biometrics** 37, 391-411.

Hoerl, A. E. and Kennard, R. W. (1970a). "Ridge regression: biased estimation for non orthogonal problems." **Technometrics** 12, 55-67.

Hoerl, A. E. and Kennard, R. W. (1970b). "Ridge regression: applications to non orthogonal problems." **Technometrics** 12, 69-82.

Golub, G. H., Heath, M., and Wahba, G. (1979). "Generalized cross-validation as a method for choosing a good ridge parameter." **Technometrics** 21, 215-223.

Legendre, A. M. (1805). **Nouvelles méthodes pour la determination des orbits des comètes.** Appendix: Sur la méthode des moindre carrés (least squares.)

Lindley, D. V. (1962). "Discussion." [of "Confidence sets for the mean of a multivariate normal distribution" by C. M. Stein.] **Journal Royal Statistical Society**, B24, 285-287.
Marquardt, D. W. and Snee, R. D. (1975). "Ridge regression in practice." **The American Statistician**, 29, 3-19.

McDonald, G. C. and Schwing, R. C. (1973). "Instabilities of regression estimates relating air pollution to mortality." **Technometrics**, 15, 463-481.

Obenchain, R. L. (1975a). "Residual optimality: ordinary vs. weighted vs. biased least squares." **Journal of the American Statistical Association**, 70, 375-379.

Obenchain, R. L. (1975b). "Ridge analysis following a preliminary test of the shrunken hypothesis." **Technometrics**, 17, 431-441. (Discussion: McDonald, G. C., 443-445.)

Obenchain, R. L. (1976). "Methods of ridge regression." **Proceedings of the Ninth International Biometric Conference**, Invited Papers, Volume One, 37-57, Boston.

Obenchain, R. (1977). "Classical F-tests and confidence regions for ridge regression." **Technometrics** 19, 429-439.

Obenchain, R. (1980). Comment on "A critique of some ridge regression methods" by G. Smith and F. Campbell. **Journal American Statistical Association** 75, 95-96.

Obenchain, R. L. (1981). "Maximum likelihood ridge regression and the shrinkage pattern hypotheses." Abstract 81t-23. **I.M.S. Bulletin** 10, 37.

Obenchain, R. L. (1984). "Maximum likelihood ridge displays." **Communications in Statistics A**, 13, 227-240. (Proceedings of the Fordham Ridge Symposium, ed. H. D. Vinod.)

Sclove, S. (1968). "Improved estimators of coefficients in linear regression." **Journal of the American Statistical Association** 63, 596-606.

Tukey, J. W. (1975). "Instead of Gauss-Markov Least Squares; What?" **Applied Statistics**, ed. R. P. Gupta. Amsterdam-New York: North Holland Publishing Company.

Further Reading for Chapter One

Casella, G. (1980). "Minimax ridge regression estimation." **Annals of Statistics** 8, 1036-1056.

Casella, G. (1985). "Condition numbers and minimax ridge-regression estimators." **Journal American Statistical Association** 80, 753-758.

Dempster, A. P., Schatzoff, M. and Wermuth, N. (1976). "A simulation study of alternatives to ordinary least squares." **Journal American Statistical Association,** 72, 77-91 (with discussion, pp. 91-106; see, especially, the discussion by Efron and Morris.)

Goldstein, M. and Smith, A. F. M. (1974). "Ridge-type estimators for regression analysis." **Journal Royal Statistical Society B**, 36, 284-291.

Hocking, R. R. (1972). "Criteria for selection of a subset regression: which one should be used?" **Technometrics**, 14, 967-970.

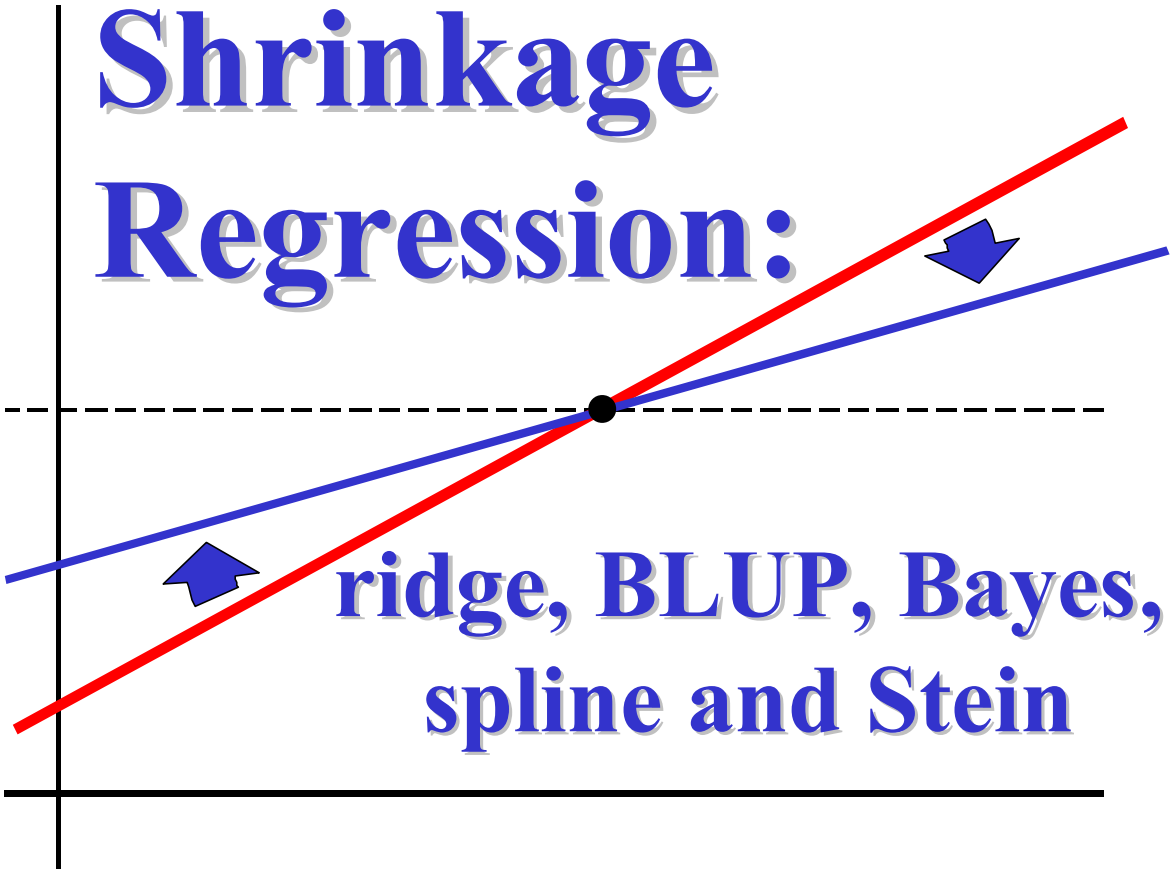
Hocking, R. R. (1976). "The analysis and selection of variables in linear regression." **Biometrics** 32, 1-49.

Lindley, D. Y. and Smith, A. F. M. (1972). "Bayes estimates for the linear model." **Journal Royal Statistical Society**, Series B, 34, 1-72.

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

Piegorsch, W. W. and Casella, G. (1989). "The early use of matrix diagonal increments in statistical problems." **Siam Review** 31, 428-434.

Theil, H. (1963). "On the use of incomplete prior information in regression analysis." **JASA** 58, 401-414.



Chapter 02: Basic Linear Model Concepts

Bob Obenchain, Ph.D.
softR$_X$ freeware
13212 Griffin Run
Carmel, Indiana 46033-8835

Chapter 2: BASIC LINEAR MODEL CONCEPTS

Multiple linear regression models quantify the relationship of a “response" variable, Y, to P given non-constant “regressor" variables (also called “predictor" or “independent" variables), $X_1, \ldots, X_P$. Wherever reasonable in this book, we will use standardized symbols and vector-matrix notation. For example, the N observed response values are formed into an N by 1 column vector, **y**, and the regressor values are formed into an N by P matrix, **X**. In other words, rows represent observations, and columns represent variables. Our linear model then becomes . . .

Conditional Expectation of **y** given **X**: $\quad E(\mathbf{y} \mid \mathbf{X}) = \mathbf{1}\,\mu + \mathbf{X}\,\boldsymbol{\beta}$ $\quad \{ 2.1 \}$

Conditional Variance of **y** given **X**: $\quad V(\mathbf{y} \mid \mathbf{X}) = \sigma^2\,\mathbf{I}$ $\quad \{ 2.2 \}$

where

$\mathbf{1}$ = column vector of N ones,
μ = unknown intercept term (scalar valued),
$\boldsymbol{\beta}$ = column vector of P unknown, true regression coefficients, and
σ^2 = unknown residual variance (non-negative scalar).

Individual regression coefficients may be visualized as being either fixed or random. A multiple linear regression model with both fixed and random coefficients is said to be a mixed model.

2.1 Centered Variables

The above linear model can be restated in terms of centered variables as follows. Suppose that the mean response value has been subtracted from each row of the response vector, so that $\mathbf{1}^T\mathbf{y} = 0$, and the row vector of regressor (column) means, $\bar{\mathbf{x}}^T$, has been subtracted from each row of the regressor matrix, so that $\mathbf{1}^T\mathbf{X} = \mathbf{0}^T$. Note that centering is equivalent to replacing the **y** vector by $(\mathbf{I} - \mathbf{1}\,\mathbf{1}^T/N)\,\mathbf{y}$ and replacing the **X** matrix by $(\mathbf{I} - \mathbf{1}\,\mathbf{1}^T/N)\,\mathbf{X}$. Our pair of model equations then become . . .

Conditional Expectation of **y** [centered] given **X** [centered] :

$$E(\mathbf{y} \mid \mathbf{X}) = \mathbf{X}\beta \qquad \{ 2.3 \}$$

Conditional Variance of **y** [centered] given **X** [centered] :

$$V(\mathbf{y} \mid \mathbf{X}) = \sigma^2 (\mathbf{I} - \mathbf{1}\mathbf{1}^T / N) \qquad \{ 2.4 \}$$

Notice the key difference between equations { 2.2 } and { 2.4 }. In equation { 2.2 }, the variance matrix, $\sigma^2 \mathbf{I}$, expresses the familiar condition that the response disturbance terms, $\mathbf{y} - E(\mathbf{y} \mid \mathbf{X})$, although possibly not statistically independent and identically distributed, are at least uncorrelated with a common variance (the so-called "homoscedastic" observations case.) The corresponding notation for the variance matrix of the centered response vector is $\sigma^2(\mathbf{I} - \mathbf{1}\mathbf{1}^T / N)$ of equation { 2.4 }, which will be much less familiar to many readers. This notation simply reminds us that the set of N deviations of response values from their common mean value, $\mathbf{y} - \bar{y}\mathbf{1}$, have the variance-covariance structure of "interchangable" (or "exchangeable") random variables. These deviations cannot be uncorrelated; the "centering" process gave them a non-random sum of ZERO.

The potential advantages of using centered-variable notation are illustrated by the following pair of formulas for the "ordinary least squares estimator," $\mathbf{b}^o$, of the P elements of β in the so-called "full rank" case. In un-centered notation,

$$\mathbf{b}^o = [\mathbf{X}^T (\mathbf{I} - \mathbf{1}\mathbf{1}^T / N) \mathbf{X}]^{-1} \mathbf{X}^T (\mathbf{I} - \mathbf{1}\mathbf{1}^T / N) \mathbf{y} ; \qquad \{ 2.5 \}$$

but, when re-expressed in terms of centered variables,

$$\mathbf{b}^o = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} . \qquad \{ 2.6 \}$$

On the other hand, it is not really necessary to assume (as in formulas { 2.5 } and { 2.6 }) that the matrix of centered sums-of-squares and cross-products, $\mathbf{X}^T(\mathbf{I} - \mathbf{1}\mathbf{1}^T / N)\mathbf{X}$, is of full rank (P) in order to define $\mathbf{b}^o$; see equation { 2.9 } below. Our point here is simply that the second expression illustrates the great gain in notational simplicity that can result from adopting the convention that all variables have been centered. Therefore, all remaining formulas in these notes are displayed tacitly assuming, unless specifically stated otherwise, that all variables have been centered by subtracting off the appropriate column mean from every row of the corresponding response vector and/or regressor matrix.

2.2 The Special Case of Uncorrelated Regressors

In the special case where the regressor variables are uncorrelated, the centered $\mathbf{X}^T\mathbf{X}$ matrix will be diagonal. And the least squares solution vector of equation { 2.6 } then yields individual coefficients of a particularly simple form. Namely, the i-th coefficient is then

$$b_i^o \;=\; \frac{\mathbf{x}_i^T\,\mathbf{y}}{\mathbf{x}_i^T\,\mathbf{x}_i} \;=\; \sqrt{\frac{\mathbf{y}^T\,\mathbf{y}}{\mathbf{x}_i^T\,\mathbf{x}_i}}\;\; r_{yx_i}\,, \qquad \{\,2.7\,\}$$

where r_{yx_i} denotes the "marginal" correlation between the response vector, $\mathbf{y}$, and the i-th regressor variable, $\mathbf{x}_i$ = i-th column of the $\mathbf{X}$ matrix.

Models with uncorrelated-regressors rarely occur in actual practice (except, possibly, in "designed" experiments), but it is interesting to note how least squares estimates are constructed in this limiting case. Note that each least squares regression coefficient is directly proportional to the corresponding marginal correlation in { 2.4 }; in particular, a fitted coefficient is guaranteed to have the same numerical sign as the marginal correlation between regressor and response coordinates. On the other hand, note that the numerical magnitude of an individual coefficient can depend as much upon your choice of scaling, $\sqrt{\mathbf{x}_i^T\mathbf{x}_i}$, for its regressor coordinates as it does upon the marginal correlation of that regressor with the response. (The scaling of the response variable, embodied by the $\sqrt{\mathbf{y}^T\mathbf{y}}$ term, applies equally to all coefficients.) Specifically, the i-th coefficient can be relatively large simply because the observed coordinates along the i-th regressor axis have relatively small "spread" (a small $\sqrt{\mathbf{x}_i^T\mathbf{x}_i}$ term) rather than because the corresponding marginal correlation is relatively large. As we will see later (in equation { 2.23 }), the statistical significance of a fitted regression coefficient depends only upon the corresponding correlation coefficient, but its numerical size can be "distorted" by any unusual, extreme choice for the scaling of its regressor/predictor coordinates.

2.3 Canonical Form of Regressors

Here, we illustrate how the basic correlation-spread relationships observed in { 2.7 } apply to the general case where regressors are usually intercorrelated. However, to see these relationships clearly in equation { 2.16 }, we will first need to "rotate" axes to canonical form. Our notation will make frequent reference to the component parts, H, Λ and G, of the singular value decomposition of the centered regressor matrix:

$$\mathbf{X} \;=\; (\,\mathbf{I} \;-\; \mathbf{1}\,\mathbf{1}^T/\,N\,)\,\mathbf{X} \;=\; \mathbf{H}\,\mathbf{\Lambda}^{1/2}\,\mathbf{G}^T, \qquad \{\,2.8\,\}$$

where

H is a semi-orthogonal (N by R) matrix of standardized "principal coordinates of **X**,"

G is a semi-orthogonal (P by R) matrix of "principal axis direction cosines for **X**,"

$\mathbf{\Lambda}^{1/2}$ is a diagonal (R by R) matrix of "ordered singular values of **X**," and

R denotes the RANK of **X**, where $1 \leq R \leq P \leq N$.

Like the centered **X** matrix, the principal coordinates matrix, **H**, has a zero mean vector, $\mathbf{1}^T\mathbf{H} = \mathbf{0}^T$. But the **H** matrix also has the properties that $\mathbf{H}^T\mathbf{H} = \mathbf{I}$ (R by R) and $\mathbf{H}\mathbf{H}^T$ is the orthogonal projection matrix onto the column space of **X**, which is a uniquely determined, symmetric, and idempotent matrix that is N by N of rank R, Rao(1973), page 47.

The principal axis direction cosine matrix, **G**, always has columns ($\vec{\mathbf{g}}_1, \ldots, \vec{\mathbf{g}}_R$) that are mutually orthogonal and of length one: $\mathbf{G}^T\mathbf{G} = \mathbf{I}$ (R by R.) In the so-called "full rank" case ($R = P$), $\mathbf{G}\mathbf{G}^T$ is also a P by P identity matrix (i.e., **G** is orthogonal rather than simply semi-orthogonal.) But, when R is strictly less than P, $\mathbf{G}\mathbf{G}^T$ is the orthogonal projection matrix for the row space of **X**, which is a uniquely determined, symmetric, and idempotent matrix of rank R that is $P \times P$.

The ordered singular values of **X** (from upper-left to lower-right along the main diagonal of $\mathbf{\Lambda}^{1/2}$) are

$$\lambda_1^{1/2} \geq \lambda_2^{1/2} \geq \ldots \lambda_R^{1/2} > 0. \qquad \{ 2.9 \}$$

Note, specifically, that each of the first R singular values is strictly positive (greater than zero.)

When **X** is NOT of full (column) rank (i.e., $R < P$), the final $P - R$ singular values of **X** are exact zeros, which were dropped from equations { 2.8 } and { 2.9 }. When $P - R \geq 2$, there would be no unique way to add 2 or more columns to the **H** and **G** matrices. Similarly, when two (or more) of the ordered singular values in { 2.9 } are exactly equal, the corresponding columns of **H** and **G** are also not uniquely determined. (It is always possible to, say, multiply all of the elements in any column of **H** by -1 if one also multiplies all elements in the corresponding column of **G** by -1. This sort of modification of the singular valve decomposition is too "trivial" to cause any real concern about "uniqueness.")

Non-uniqueness caused by one or more zero singular values carries over to estimates of the $\boldsymbol{\beta}$ coefficient vector, at least when estimates are viewed as "solutions" to under-determined "normal equations," $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y}$. Twenty years ago, $\boldsymbol{\beta}$ was commonly said to be "not estimable" whenever $R < P$. Today, it is apparently more common to adopt the convention that "the" least squares estimator of $\boldsymbol{\beta}$ is always (even in the less-than-full-rank-case) defined to be...

$$\mathbf{b}^o \equiv \mathbf{X}^+\mathbf{y}\,, \qquad \{ 2.10 \}$$

where the $+$ superscript denotes the (unique) Moore-Penrose inverse of the (centered) regressor matrix, **X** , Rao(1973), page 26. It is straightforward to show, using { 2.8 } and { 2.9

}, that $\mathbf{X}^{+} = \mathbf{G}\,\mathbf{\Lambda}^{-1/2}\,\mathbf{H}^{\mathrm{T}}$ when R is either less than or equal to P. Other implications of this factorization are explored in great detail below.

NUMERICAL EXAMPLE:

Table 2.1 below lists the data for a small numerical example with $N = 10$ observations on $P = 2$ regressor variables and a response variable.

Table 2.1 A Small Numerical Example

X1	X2	Y
− 1.67	− 1.68	− 1.58
− 1.11	− 0.34	− 1.06
− 0.58	− 1.35	− 0.53
− 0.28	− 0.21	− 0.79
− 0.54	0.00	− 0.48
0.28	− 0.34	0.74
0.56	0.99	1.06
0.84	0.67	0.79
1.11	1.35	0.53
1.39	0.91	1.32

Besides being centered as explained in §2.1, so that the mean value in each column is 0, the numerical values of each variable have also been placed on a standardized scale. This rescaling involves dividing each original predictor coordinate by the sample standard deviation of that variable. As a result, the sum-of-squares of the $N = 10$ values in each column of Table 2.1 is $N - 1 = 9$. And the pairwise sample correlations between variables are

	X_1	X_2	Y
X_1	1.0000		
X_2	0.8683	1.0000	
Y	0.9401	0.7901	1.0000 .

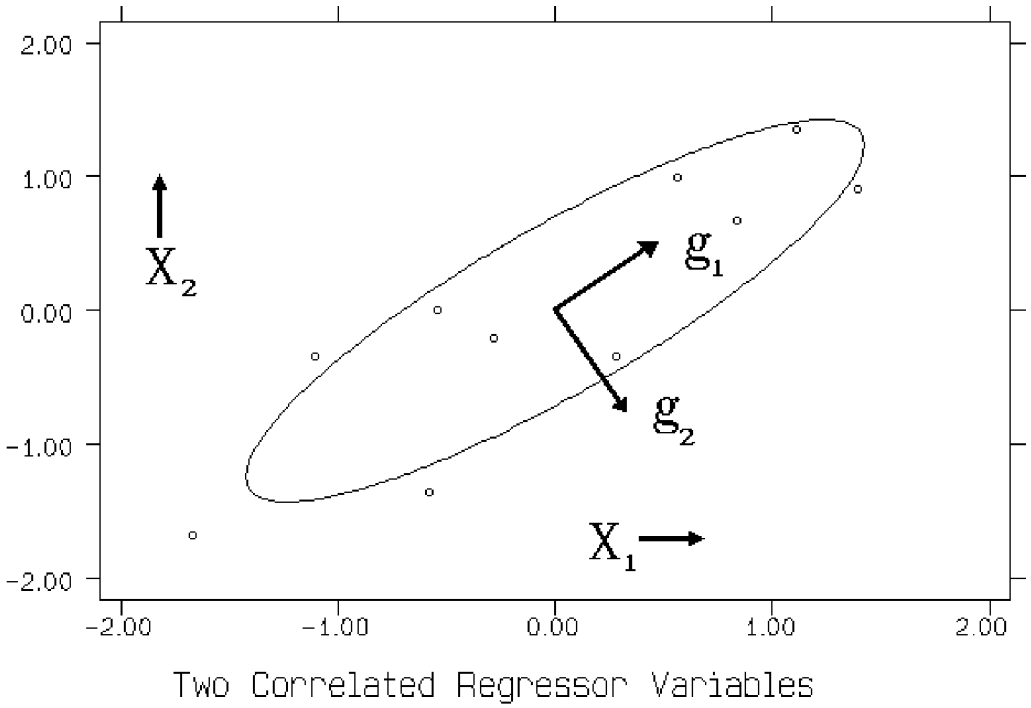
Because regressor variables have exactly equal variance following this rescaling, principal axes for the bivariate ($P = R = 2$) case will always be rotated to exactly a 45° angle relative to the given regressor axes. Although not uniquely determined, each element of the 2×2 matrix, $\mathbf{G}$, of principal axis direction cosines will be $\pm \sqrt{2}$. For example, one possible choices is

$$G = \begin{bmatrix} 0.707 & 0.707 \\ 0.707 & -0.707 \end{bmatrix}.$$

By the way, rescaling regressors to have equal variance when $P \geq 3$ does not necessarily yield a **G** matrix of any special form (like it does here in the bivariate case, $P = 2$.)

Figure 2.1 displays a scatter-plot of the X_1 and X_2 regressor coordinates from Table 2.1 and the $\vec{g}_1$ and $\vec{g}_2$ axes corresponding to the above choice for the columns of **G** . Furthermore, a constant-density ellipse of a hypothetical bivariate-normal distribution for a pair of standardized variables with the same inter-correlation, $\widehat{\rho}_{12} = +0.8683$, as that observed between X_1 and X_2 is also shown in Figure 2.1.

Figure 2.1 Given Regressor Coordinates and Principal Axes



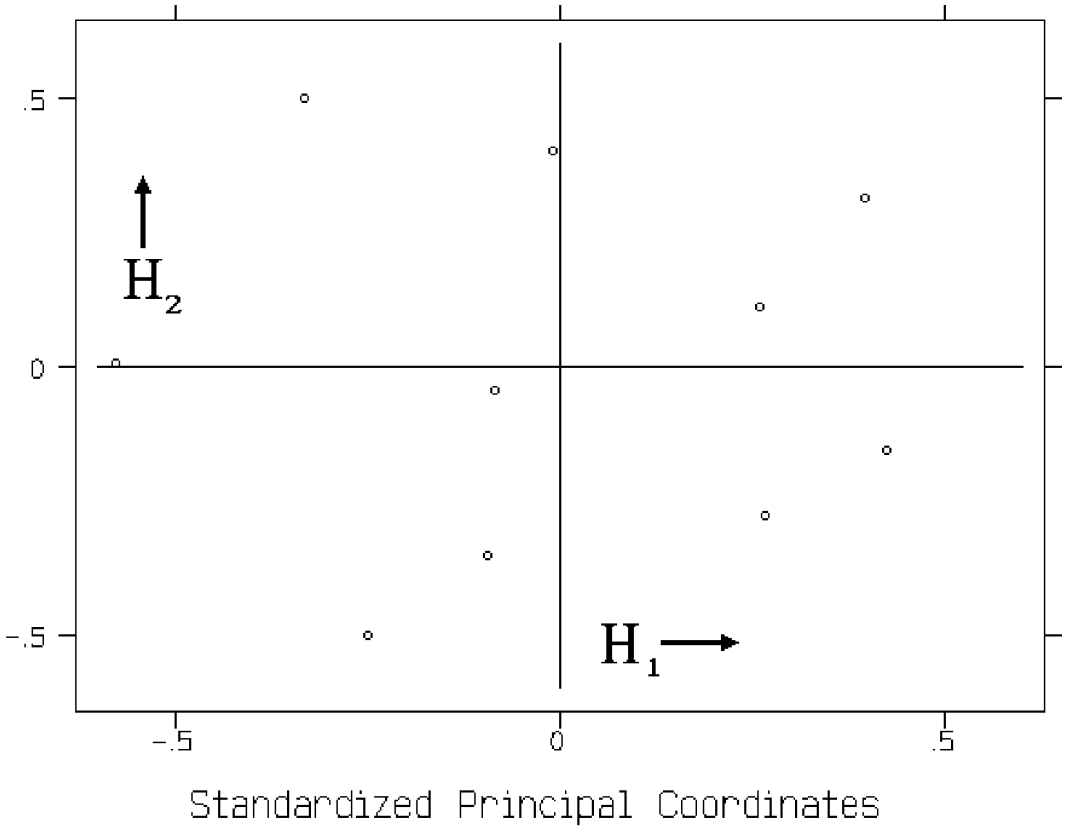
The singular values of the **X** matrix are $\lambda_1^{1/2} = \sqrt{(N-1)\cdot(1+\widehat{\rho}_{12})} = 4.1006$ and $\lambda_2^{1/2} = \sqrt{(N-1)\cdot(1-\widehat{\rho}_{12})} = 1.0886$. And the 10×2 **H** matrix of standardized regressor principal coordinates is

$$\mathbf{H} = \begin{matrix} -0.5778 & 0.0059 \\ -0.2502 & -0.5006 \\ -0.3329 & 0.4999 \\ -0.0845 & -0.0456 \\ -0.0932 & -0.3510 \\ -0.0103 & 0.4028 \\ 0.2673 & -0.2791 \\ 0.2605 & 0.1107 \end{matrix}$$

0.4243 – 0.1555
0.3968 0.3123

Figure 2.2 displays a scatter-plot of these standardized H_1 and H_2 coordinates, showing that they are uncorrelated with equal variance. (The sum-of-squares of the N = 10 values in each column of the **H** matrix is 1.)

Figure 2.2 Regressor Principal Coordinates



2.4 Numerical versus Statistical Ill-Conditioning

Terminology: A multiple regression problem is called either EXACTLY SINGULAR or NUMERICALLY ILL-CONDITIONED whenever the matrix of centered regressor coordinates is less than full rank.

In actual regression practice, exact singularities and/or ties among singular values are rather rare (except, again, in designed “orthogonal" experiments.) Anyway, regression practitioners usually find themselves in the common situation where the singular values of X are distinct...

$$\lambda_1^{1/2} > \lambda_2^{1/2} > \ldots > \lambda_P^{1/2} > 0 . \qquad \{ 2.11 \}$$

All of the component parts [$\mathbf{H}$, $\mathbf{\Lambda}$ and $\mathbf{G}$] of the singular value decomposition are “essentially" uniquely determined when { 2.11 } holds.

Statistical ill-conditioning is a much more common problem for regression practitioners than is numerical ill-conditioning. In fact, statistical ill-conditioning is almost always present in at least some very weak form whenever data collection had failed to follow a well-planned design. We can give the following “qualitative" definition now; much greater insight into this general topic will be provided below in Section §2.6.

A multiple regression problem is said to be STATISTICALLY ILL-CONDITIONED when the trailing singular values, $\lambda_P^{1/2}$, $\lambda_{P-1}^{1/2}, \ldots$, of the centered regressor matrix, $\mathbf{X}$, are numerically small compared to its leading singular values, $\lambda_1^{1/2}$, $\lambda_2^{1/2}, \ldots$

2.5 Eigen Decompositions

The familiar eigenvalue-eigenvector decomposition of the regressor adjusted sums-of-squares and cross-products matrix, $\mathbf{X}^T \mathbf{X}$, is closely related to the singular value decomposition of equation { 2.7 } in the sense that:

$$\mathbf{X}^T\mathbf{X} = \mathbf{X}^T(\mathbf{I} - \mathbf{1}\mathbf{1}^T/ N)\mathbf{X} = \mathbf{G}\mathbf{\Lambda}\mathbf{G}^T. \qquad \{ 2.12 \}$$

This well-known decomposition technique is useful in numerical computations where N is much, much larger than P. Rather than attack the N by P centered regressor matrix, $\mathbf{X}$, via the singular value decomposition when N >> P, one can use the eigenvalue-eigenvector approach on a much smaller (P by P) matrix to calculate only $\mathbf{\Lambda}$ and $\mathbf{G}$. Later, when actually needed, principal coordinates can always be computed (if only approximately) by simple matrix multiplication, $\mathbf{H} = \mathbf{X}\,\mathbf{G}\,\mathbf{\Lambda}^{-1/2}$.

The ordered eigenvalues of $\mathbf{X}^T \mathbf{X}$, which are denoted by

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_R > 0, \qquad \{ 2.13 \}$$

are the "adjusted" (via centering) sums-of-squares of **X** coordinates along their principal axes. In other words, an eigenvalue is (N – 1) times the sample variance along a certain direction in P-dimensional regressor space. Similarly, the corresponding singular value (square root of the eigenvalue) is $\sqrt{N-1}$ times a sample standard deviation along that same direction in regressor space.

2.6 The Uncorrelated Components of Least Squares

The notation and terminology introduced above allows us to examine, in great detail, the basic structure of the least squares estimator of $\boldsymbol{\beta}$. Specifically, we can rewrite equation { 2.9 } as

$$\mathbf{b}^o = \mathbf{G}\,\mathbf{c}\,, \qquad \{ 2.14 \}$$

where **G** is the semi-orthogonal direction cosines matrix for the principal axes of the centered regressors and **c** is the R × 1 column vector containing the uncorrelated components of $\mathbf{b}^o$. Thus, by definition, the vector of uncorrelated components is of the form:

$$\begin{aligned} \mathbf{c} &\equiv \mathbf{G}^T \mathbf{b}^o, \\ &= \boldsymbol{\Lambda}^{-1/2}\,\mathbf{H}^T\,\mathbf{y}\,, && \{ 2.15 \} \\ &= \sqrt{\mathbf{y}^T\mathbf{y}}\cdot\boldsymbol{\Lambda}^{-1/2}\,\mathbf{r}\,. && \{ 2.16 \} \end{aligned}$$

In equation { 2.16 } , $\mathbf{r} = (r_{yi})$ represents the column vector of principal correlations between the response vector, **y**, and the columns of the principal axis regressor coordinate matrix, **H**. Specifically, $\mathbf{r} = \mathbf{H}^T\,\mathbf{y} / \sqrt{\mathbf{y}^T\mathbf{y}}$, which is an R × 1 vector. As is stressed below, equation { 2.16 } has a lot to say about the basic nature of statistical ill-conditioning in multiple linear regression models!

However, we first observe that $\mathbf{b}^o$ will be an unbiased estimator of $\boldsymbol{\beta}$ when rank(**X**) = R = P as long as the expectation models, { 2.1 } and { 2.3 }, are "correct" statistical models for the data at hand. In other words, whenever β is estimable we have

$$E (\mathbf{b}^o \mid X) = \mathbf{X}^+\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}\,. \qquad \{ 2.17 \}$$

Furthermore, if the variance models, { 2.2 } and { 2.4 }, are "correct" models, the variance matrix of $\mathbf{b}^o$ will be of the form:

$$V (\mathbf{b}^o \mid \mathbf{X}) = \sigma^2\,\mathbf{X}^+\,\mathbf{X}^{+T} = \sigma^2\,(\mathbf{X}^T\,\mathbf{X})^+. \qquad \{ 2.18 \}$$

Of course, equation { 2.18 } becomes simply $V(\mathbf{b}^o \mid \mathbf{X}) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$ when $R = \text{rank}(\mathbf{X}) = P$.

The corresponding expectation vector and variance matrix for the uncorrelated components are

$$E(\mathbf{c} \mid \mathbf{X}) \equiv \gamma = \mathbf{G}^T \beta , \qquad \{ 2.19 \}$$

and

$$V(\mathbf{c} \mid \mathbf{X}) = \sigma^2 \mathbf{\Lambda}^{-1} . \text{ (R by R)} \qquad \{ 2.20 \}$$

Note, specifically, that our terminology here is motivated by the observation that the elements of **c** are always uncorrelated (i.e. their variance matrix is always diagonal.)

Developing an appreciation for the implications of formula { 2.16 } is fundamental to understanding ill-conditioning as a dual numerical/statistical phenomenon. Therefore, let us now examine the individual elements of the uncorrelated component **c** vector. The i-th such element is...

$$c_i = \sqrt{\mathbf{y}^T \mathbf{y}} \cdot r_{yi} / \lambda_i^{1/2} . \qquad \{ 2.21 \}$$

Individual uncorrelated components can be relatively large, <u>numerically</u>, either because their corresponding principal correlation is relatively large or simply because their corresponding regressor singular value is relatively small.

On the other hand, note from equation { 2.20 } that the standard error of the i-th uncorrelated component, c_i , is σ divided by $\lambda_i^{1/2}$. In other words, the "relative" standard error of c_i is the known quantity $\lambda_i^{-1/2}$. Therefore, one's uncertainty about the size of each true component of β is inversely proportional to the spread in regressor coordinates along the corresponding principal axis.

NUMERICAL EXAMPLE: Let us now continue the numerical example ($P = 2$, $N = 10$) we introduced in section §2.3. The marginal correlations between the response **y** values and the two given regressor **X** columns were previously stated to be $\hat{\rho}_{y1} = +0.9401$ and $\hat{\rho}_{y2} = +0.7901$. The uncorrelated components of $\mathbf{b}^o$ are thus $c_1 = +0.654828$ and $c_2 = +0.805537$ in { 2.15 }, and the principal correlations of the response **y** values with the two sets of regressor principal **H** coordinates are $r_{y1} = +0.8951$ and $r_{y2} = +0.2923$ in { 2.16 }. Note that r_{y1} is more than 3 times larger than r_{y2} . Yet c_1 is smaller, numerically, than is c_2 . And we know why! The second component, c_2 , is greatly magnified because the singular value in its denominator, $\lambda_2^{1/2} = 1.089$, is much smaller than the singular value, $\lambda_1^{1/2} = 4.101$, in the denominator of c_1 .

2.7 Statistical Significance of Uncorrelated Components

The F-ratio for testing the hypothesis that $\gamma_i = 0$ (i.e. testing the "statistical significance" of the i-th estimated component, c_i) is of the form:

$$F_i = \frac{c_i^2 \, \lambda_i}{s^2} . \qquad \{ 2.22 \}$$

where the least-squares residual-mean-square, $s^2 = \mathbf{y}^T (\mathbf{I} - \mathbf{H} \mathbf{H}^T) \mathbf{y} / (N - R - 1)$, is the estimator of the error variance, σ^2.

But, wait a second! A little bit of algebra reveals that the F-statistic of { 2.22 } does not actually depend upon the λ_i or any of the other regressor eigenvalues or singular values! Using expression { 2.22 } for c_i, an expression equivalent to { 2.22 } is:

$$F_i = \frac{(N-R-1) \cdot r_{yi}^2}{(1 - R^2)} \qquad \{ 2.23 \}$$

where R is the rank of $\mathbf{X}$ and R^2 is the familiar R-squared statistic, which can be expressed as the following sum-of-squares of principal correlations:

$$R^2 = r_{y1}^2 + r_{y2}^2 + \ldots + r_{yR}^2 .$$

The t-statistic corresponding to { 2.23 } is

$$t_i = r_{yi} \cdot \sqrt{\frac{(N-R-1)}{(1 - R^2)}} . \qquad \{ 2.24 \}$$

Equations { 2.23 } and { 2.24 } remind us that . . .

Individual uncorrelated components CANNOT be judged relatively important, statistically, simply because they are large numerically. A component can be large simply because its corresponding regressor singular value is relatively small. Statistical significance depends only upon the regressor principal correlations.

NUMERICAL EXAMPLE:

Again, continuing the numerical example (P = 2, N = 10) from sections §2.3 and §2.6 , the R-squared statistic is $R^2 = 0.8951^2 + 0.2923^2 = 0.8866$. And the t-statistics for the individual uncorrelated components, 7.03 and 2.30 , are directly proportional to the corresponding principal correlations, 0.8951 and 0.2923 .

2.8 Predictions, Residuals and Linear Reparameterizations that remove Ill-Conditioning.

Like the above F-ratios and t-statistics, it is easily shown that least-squares predicted responses and residuals also do not depend upon theregressor eigenvalues or singular values, $\mathbf{\Lambda}$ of { 2.8 }. In fact, these quantities also do not depend upon the direction cosines, **G** of { 2.8 }. Specifically, the vector of [centered] least-squares predictions is

$$\mathbf{y}^o = \mathbf{X}\,\mathbf{b}^o = \mathbf{X}\,\mathbf{X}^+\mathbf{y} = \mathbf{H}\mathbf{\Lambda}^{1/2}\mathbf{G}^T\mathbf{G}\mathbf{\Lambda}^{-1/2}\mathbf{H}^T\mathbf{y} = \mathbf{H}\,\mathbf{H}^T\mathbf{y}\,,$$

where $\mathbf{H}\,\mathbf{H}^T$ is again the uniquely determined orthogonal projection matrix for the R-dimensional column space of the centered **X** matrix. The corresponding vector of [centered] residuals is

$$\mathbf{y} - \mathbf{y}^o = (\mathbf{I} - \mathbf{H}\,\mathbf{H}^T)\mathbf{y} = (\mathbf{I} - \mathbf{1}\mathbf{1}^T/N - \mathbf{H}\,\mathbf{H}^T)\mathbf{y}\,.$$

These residuals are used to define s^2, the estimate of σ^2 used in equation { 2.22 }. Note that $(\mathbf{I} - \mathbf{1}\mathbf{1}^T/N - \mathbf{H}\,\mathbf{H}^T)$ is the uniquely determined orthogonal projection matrix for the (N – R – 1) dimensional linear subspace orthogonal to both **1** and the column space of the centered **X** matrix.

Although it is true that $\mathbf{H} = \mathbf{X}\mathbf{G}\mathbf{\Lambda}^{-1/2}$, this relationship does NOT establish that the **H** matrix of standardized principal coordinates actually depends upon either **G** or $\mathbf{\Lambda}$ (at least in the usual situation where all of the singular values of **X** are distinct.) Rather, our point-of-view is that the singular value decomposition of **X** in equation { 2.8 } simultaneously defines the **H, G** and $\mathbf{\Lambda}$ matrices. And the above arguments describe senses in which the **G** and $\mathbf{\Lambda}$ matrices can be "disregarded," leaving only the standardized **H** coordinates to define least squares predictions for both the response vector and the residual vector

In view of the above observations, let us now consider the following linear reparameterization of our original, ill-conditioned regression model, { 2.3 } and { 2.4 }. Specifically, the centered **X** matrix is now replaced by its semi-orthogonal (N by R) matrix of principal coordinates, **H** = **XA,** where the linear transformation matrix is $\mathbf{A} = \mathbf{G}\mathbf{\Lambda}^{-1/2}$.

Conditional Expectation of **y** [centered] given **H** [centered] :

$$E(\mathbf{y} \mid \mathbf{H}) = \mathbf{H}\boldsymbol{\alpha} \qquad \{ 2.25a \}$$

Conditional Variance of **y** [centered] given **H** [centered] :

$$V(\mathbf{y} \mid \mathbf{H}) = \sigma^2 (\mathbf{I} - \mathbf{1}\mathbf{1}^T / N) \qquad \{ 2.26a \}$$

We now argue that this reparameterized regression model displays absolutely no ill-conditioning (numerical or statistical) in the senses discussed above. First, note that $\boldsymbol{\alpha}$ is $R \times 1$ like $\boldsymbol{\gamma}$ rather than $P \times 1$ like $\boldsymbol{\beta}$. Furthermore, the implied **G** and $\boldsymbol{\Lambda}$ matrices for this reparameterized regression can both be taken to be $R \times R$ identity matrices. And the least squares estimate of $\boldsymbol{\alpha}$ in { 2.3a } is $\mathbf{a}^o = \mathbf{H}^+\mathbf{y} = \mathbf{H}^T\mathbf{y}$. Thus the i-th element of $\mathbf{a}^o$ is of the form

$$a_i^o = \sqrt{\mathbf{y}^T\mathbf{y}} \cdot r_{yi} = c_i \lambda_i^{1/2}, \qquad \{ 2.27a \}$$

and it follows that the elements of $\mathbf{a}^o$ are uncorrelated and homoscedastic (i.e. have a common variance of σ^2.)

Although completely free of ill-conditioning in the above senses, estimates for this reparameterized model are, none the less, closely related to corresponding estimates for the original, ill-conditioned model. Equation { 2.21a } shows that least squares estimates for regression coefficients differ only by a known scale factor. Furthermore, both models yield the exact same response predictions and residuals! Finally, the F-ratio or t-statistic for testing the significance of an individual $\boldsymbol{\alpha}$ coefficient estimate is identical to the F-ratio or t-statistic for testing the significance of the corresponding element of the least squares estimate of $\boldsymbol{\gamma}$ in the original model.

Ok, so what are some of the potential implications of the above observations?

2.8.1 Almost Irrelevant Information

My personal opinion is that the above observations show that estimation methodology designed to treat ill-conditioning in multiple regression models should have little or no effect on predicted responses or fitted residuals. Least squares is adequate (perhaps, even ideal) for these estimation tasks. The fact that the given regression parameterization is ill-conditioned is almost irrelevant when attention is restricted to only response predictions and residuals.

Methods designed to treat ill-conditioning should focus almost exclusively on problems associated with the high intercorrelations between least squares estimates of the elements of the $\boldsymbol{\beta}$ vector and/or the corresponding wildly heteroscedastic estimates for the uncorrelated components, $\boldsymbol{\gamma}$. After all, it is the given ill-conditioned **X** matrix that contains the regressor variables of genuine interest to you. You wanted to estimate their $\boldsymbol{\beta}$ coefficients in the first place because these are the variables available to you that may determine expected response, $E(\mathbf{y} \mid \mathbf{X})$. Your bad luck is that these **X** variables are highly intercorrelated in the only available

data. Furthermore, your interest in the true α coefficients corresponding to the $\mathbf{H} = \mathbf{XG\Lambda}^{-1/2}$ reparameterization is limited because this linear transformation of regressor variables is sufficiently complicated that you cannot easily visualize what it "means."

2.8.2 The Inverse Linear Transformation Restriction

Smith and Campbell(1980) argued, among other things, that the presence or absence of ill-conditioning in a multiple regression model should essentially be ignored. Specifically, they considered pairs of regressor variable reparameterizations, $\mathbf{X_1}$ and $\mathbf{X_2}$, that differ only due to an invertible linear transformation, $\mathbf{A}$, of the form $\mathbf{X_1 A = X_2}$. Elementary matrix manipulations then establish that $\mathbf{X_1}\boldsymbol{\beta_1} = \mathbf{X_1 A A^{-1}}\boldsymbol{\beta_1} = \mathbf{X_2}\boldsymbol{\beta_2}$. In other words, when a linear regression model $E(\mathbf{y}|\mathbf{X_1}) = \mathbf{X_1}\boldsymbol{\beta_1}$ is reparameterized to $E(\mathbf{y}|\mathbf{X}_2) = \mathbf{X}_2\boldsymbol{\beta}_2$, it is clear that the transformed regression coefficients must necessarily satisfy the "inverse transformation restriction" that $\boldsymbol{\beta}_2 = \mathbf{A}^{-1}\boldsymbol{\beta_1}$.

Because true regression coefficients always follow this restriction upon linear reparameterization, Smith and Campbell(1980) took the argument one step further by reasoning that statistical methodology "should" provide coefficient estimates that also satisfy this restriction. Least squares estimates do indeed always satisfy $\mathbf{b}_2^o = \mathbf{A}^{-1}\mathbf{b}_1^o$ following invertible linear regressor transformations of the form $\mathbf{X_1 A = X_2}$. Equation { 2.21a } illustrates this for the special case where the $\mathbf{A}$ reparameterization matrix is $\mathbf{G\Lambda}^{-1/2}$ and $\mathbf{A}^{-1} = \mathbf{\Lambda}^{+1/2}\mathbf{G}^{\mathrm{T}}$.

As we shall see in Chapter 3 on shrinkage regression fundamentals, "optimally" shrunken estimates generally do NOT satisfy this reparameterization restriction. This is the case because there is a non-linear relationship between the form and extent of ill-conditioning in a given parameterization and the extent of shrinkage that minimizes risk (expected or mean squared error loss) via explicit variance-bias trade-offs.

2.9 Signal-to-Noise Ratios

Perhaps there is a final “irony" here. The unknown, true non-centrality of the i-th variance ratio, F_i, is

$$\phi_i^2 \;=\; \gamma_i^2\,\lambda_i \,/\, \sigma^2 \;=\; \alpha_i^2 / \,\sigma^2. \qquad \{\ 2.28\ \}$$

Now ϕ_i^2 is a signal-to-noise ratio that plays a pivotal role in the theory of mean-squared-error optimal shrinkage along the i-th principal regressor axis. The statistical “power" parameters that control detection sensitivity for the true components of $\boldsymbol{\beta}$ are directly proportional to the spreads, the λ's, in regressor coordinates along their principal axes. Minimal spread therefore implies minimal power.

Deliberately DESIGNING a statistically ill-conditioned EXPERIMENT is almost surely unwise. These notes will concentrate upon methods useful in OBSERVATIONAL STUDIES where the regression practitioner has no hope of "controlling" or "guiding" the data collection process.

2.9 The Statistical Distribution of Principal Correlations

Even under "normal distribution theory," the exact statistical distribution of the principal correlations, $r_{y1}, \ldots, r_{yP}$, is usually not that of classical correlation coefficients. This is the case because the distribution theory of interest to us will almost always be conditional on the observed regressor values. From this point-of-view, the vector of principal correlations, $\mathbf{r} = \mathbf{H}^T \mathbf{y} / \sqrt{\mathbf{y}^T \mathbf{y}}$, is simply a known linear transformation (defined by the $\mathbf{H}^T$ matrix) of the version of the response vector, $\mathbf{y}$, that has been "rescaled" to be of unit length. The linear ($\mathbf{H}^T$) part of this transformation reduces dimensionality from N-dimensional response space down to R-dimensions. Therefore, if response values start out having a joint (multivariate) normal distribution given $\mathbf{X}$, the principal correlations end up having a "rescaled" normal distribution confined to an R-dimensional hypersphere, $\mathbf{r}^T \mathbf{r} \leq 1$, of unit radius. The principal correlations have equal variances, namely

$$V(r_{yi}) = \frac{(1-R^2)}{(N-R-1)} = \frac{s^2}{y^T y} \qquad \{ 2.29 \}$$

In our ($P = 2$, $N = 10$) numerical example, the two principal correlations each have standard error $\sqrt{(1 - 0.8866)/7} = 0.127291$.

2.10 When "Should" Coefficients have "Wrong" Signs?

Cases where a fitted least-squares regression coefficient has the "wrong" (unbelievable) numerical sign seem to arise most frequently in applications with "many" or, at least, "several" predictor (X) variables, as in the four-predictor gasoline-mileage example of Section §1.3. But this phenomenon can also be illustrated when there are only two predictors. Here in Section §2.10, we will explore only this most simple case ($P = R = 2$) of multiple regression. We start out with the usual theoretical model under which all three variables (Y, X_1 and X_2) have a joint, stochastic distribution. But we will end up relating our observations back to the common applications (of conditional inference) in which predictor coordinates are viewed as given.

NUMERICAL EXAMPLE:

Again, continuing the numerical example ($P = 2,\ N = 10$) we started in section §2.3 , the ordinary least-squares regression coefficients estimates are 1.03263 and -0.106568 . Thus the second coefficient does have the "wrong" sign in the sense we will be considering here. Because the two coefficients have equal variances in this case, the corresponding t-statistics are 4.02363 and -0.415237; thus, the second coefficient is not significantly different from zero.

2.10.1 Stochastic Response and Predictor Variables

Suppose that the joint distribution of Y, X_1 and X_2 has the vector of expected values

$$E \begin{bmatrix} Y \\ X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} \mu_y \\ \mu_1 \\ \mu_2 \end{bmatrix}, \qquad \{ 2.30 \}$$

and the matrix of variances

$$V \begin{bmatrix} Y \\ X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} \sigma_y^2 & \rho_{y1}\sigma_y\sigma_1 & \rho_{y2}\sigma_y\sigma_2 \\ \rho_{y1}\sigma_y\sigma_1 & \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{y2}\sigma_y\sigma_2 & \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}, \qquad \{ 2.31 \}$$

where each μ term represents the mean value of a variable; each σ term represents the standard deviation (square root of the variance) of a variable; each ρ term represents a correlation between two variables; and the subscripts y, 1, and 2 correspond to the variables Y, X_1, and X_2 , respectively.

When a two-predictor regression model is to be fit to data, the elements of the E vector and the V matrix shown in equations { 2.27 } and { 2.28 } are simply replaced by their natural estimates, $\widehat{\mu}_j$, $\widehat{\sigma}_j$, and $\widehat{\rho}_{jk}$ for $j = y, 1, 2$ and $k \neq j$.

The regression of Y onto a single predictor, X_j , has slope coefficient

$$\beta_j^{(1)} = E[\ (X_j - \mu_j)^2\ ]^{-1} \cdot E[\ (X_j - \mu_j) \cdot (Y - \mu_y)\ ] = \rho_{yj}\ \sigma_y / \sigma_j\ , \qquad \{ 2.32 \}$$

for $j = 1$ or 2.

PREDICTOR SIGNS CONVENTION: Note that we can change the numerical sign of X_1 and/or of X_2 , if necessary, so that neither predictor-response marginal correlation is negative: $\rho_{y1} \geq 0$ and $\rho_{y2} \geq 0$. Thus, without loss of generality, we need only consider the case where both single-regressor coefficients are non-negative in { 2.29 }: $\beta_1^{(1)} \geq 0$ and $\beta_2^{(1)} \geq 0$.

When the response variable, Y, is regressed onto both X_1 and X_2, the resulting slope coefficients are defined by the matrix equations

$$\beta = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}^{-1} \begin{bmatrix} \rho_{y1}\sigma_y\sigma_1 \\ \rho_{y2}\sigma_y\sigma_2 \end{bmatrix},$$

which yield the coefficients

$$\beta_1 = (\rho_{y1} - \rho_{12}\rho_{y2}) \sigma_y / [\sigma_1 (1 - \rho_{12}^2)], \qquad \{ 2.33 \}$$

and

$$\beta_2 = (\rho_{y2} - \rho_{12}\rho_{y1}) \sigma_y / [\sigma_2 (1 - \rho_{12}^2)], \qquad \{ 2.34 \}$$

assuming that $|\rho_{12}|$ is strictly less than 1 (i.e. assuming the inverse matrix exists!)

WRONG SIGN PROBLEMS: We will say that a "wrong sign" problem has NOT occurred as long as both $\beta_1 \geq 0$ and $\beta_2 \geq 0$ when $\rho_{y1} \geq 0$ and $\rho_{y2} \geq 0$.

Note, from equations { 2.30 } and { 2.31 }, that "wrong sign" problems cannot occur when the predictor intercorrelation is non-positive: $\rho_{12} \leq 0$. Alas, our convention of choosing the numerical signs of X_1 and X_2 so that $\beta_1^{(1)} \geq 0$ and $\beta_2^{(1)} \geq 0$, tends (at least in my experience) to yield a numerical value for ρ_{12} that is strictly positive.

We now argue that "wrong sign" problems tend to occur when $\rho_{12} > 0$ because ρ_{y1} and ρ_{y2} are spread apart, numerically. And the numerical difference between ρ_{y1} and ρ_{y2} does NOT need to be very large at all before a "wrong sign" may be produced, at least when ρ_{12} is near its upper limit of one. For example, equations { 2.30 } and { 2.31 } show that, whenever $\rho_{12} > 0$ and $\rho_{y1} > \rho_{y2}$, then

(i) β_1 will always have a "believably" positive sign, while

(ii) β_2 may be "unbelievably" negative, and certainly will be if ρ_{12} is sufficiently close to 1.

On the other hand, the exact reverse sort of situation ($\beta_1 < 0$ and $\beta_2 > 0$) can occur when $\rho_{12} > 0$ and $\rho_{y1} < \rho_{y2}$.

To consider all possibilities, we now denote the ratio of ρ_{y1} to ρ_{y2} by $\Re = \rho_{y1} / \rho_{y2}$ and rewrite { 2.30 } and { 2.31 } as

$$\beta_1 = \sigma_y \cdot \rho_{y2} \cdot (\Re - \rho_{12}) / [\sigma_1 (1 - \rho_{12}^2)]$$

and

$$\beta_2 = \sigma_y \cdot \rho_{y2} \cdot (1 - \rho_{12} \cdot \Re) / [\sigma_2 (1 - \rho_{12}^2)] .$$

Note that the numerical sign of β_1 will agree with that of $(\Re - \rho_{12})$, while the numerical sign of β_2 will agree with that of $(1 - \rho_{12} \cdot \Re)$.

For the joint distribution of Y, X_1, and X_2 be non-singular, the determinant of the **V** matrix of equation { 2.28 } must be strictly positive. This condition can be written as $\sigma_y^2 \cdot \sigma_1^2 \cdot \sigma_2^2 \cdot (1 + 2 \cdot \rho_{y1} \cdot \rho_{y2} \cdot \rho_{12} - \rho_{y1}^2 - \rho_{y2}^2 - \rho_{12}^2) > 0$. Assuming that σ_y^2, σ_1^2 and σ_2^2 are each positive, this non-singularity condition can then be rewritten in terms of the three parameters ρ_{12}, $\Re = \rho_{y1} / \rho_{y2}$ and ρ_{y2} as...

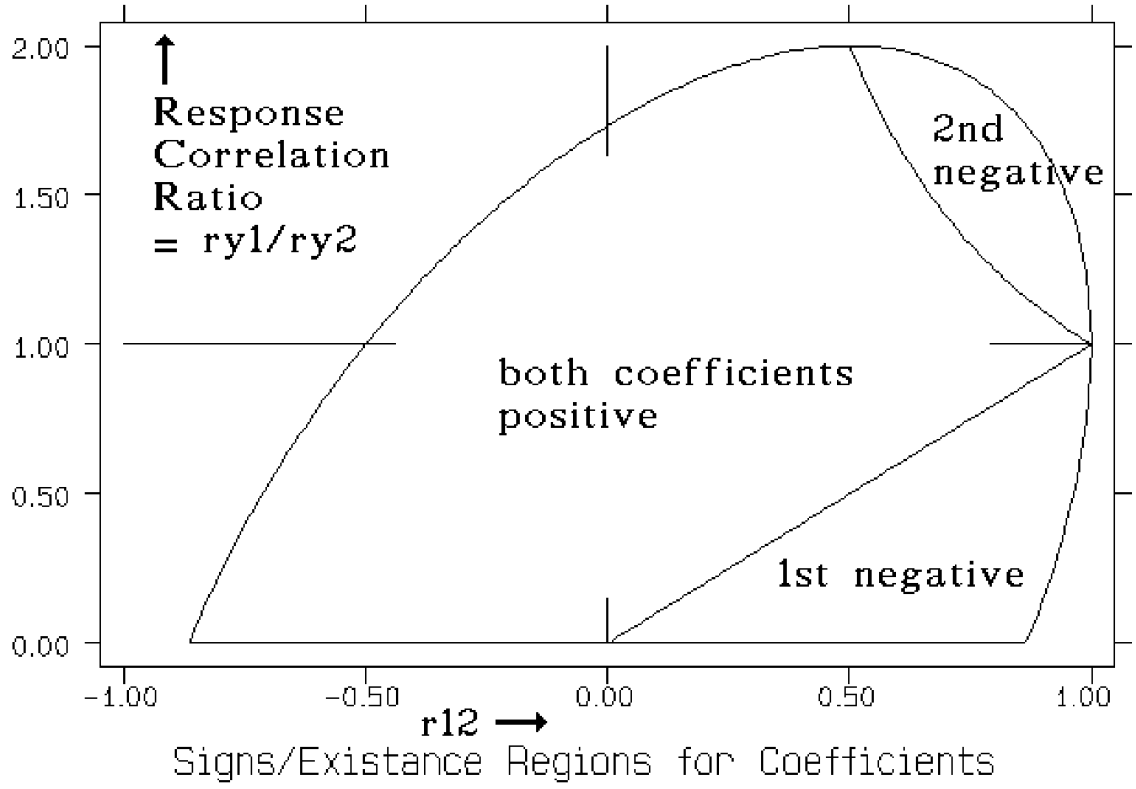
$$\max[0, \ \rho_{12} - \sqrt{(1 - \rho_{12}^2) \cdot (\frac{1}{\rho_{y2}^2} - 1)}] < \Re < \rho_{12} + \sqrt{(1 - \rho_{12}^2) \cdot (\frac{1}{\rho_{y2}^2} - 1)}$$

Figure 2.3 below illustrates how the numerical signs of β_1 and β_2 will vary, all in the special case where $\rho_{y2} = 0.5$, over subsets of the rectangular region defined by $-1 < \rho_{12} < +1$ (along the horizontal axis) and $0 \leq \Re < 1 / \rho_{y2}$ (along the vertical axis). Values outside of the upper part of the ellipse shown in Figure 2.3 are impossible in the sense that they would either violate the above tri-variate non-singularity condition or else violate the $\rho_{y1} \geq 0$ and $\rho_{y2} \geq 0$ sign convention that we adopted above.

The primary message of Figure 2.3 is that there are well defined regions in which multiple regression coefficients “should" have the “wrong" numerical sign when $\rho_{y2} = 0.5$. Of course, corresponding regions where regression coefficients either have “wrong" signs or are “undefined" also exist for all other values of ρ_{y2} over the range $0 \leq \rho_{y2} \leq 1$.

In addition to the “wrong signs" problems that can occur when $\rho_{12} > 0$, other sorts of problems are almost guaranteed to occur whenever $|\rho_{12}|$ is close to 1. In particular, note that $(1 - \rho_{12}^2)$ terms appear in the denominators of both equations { 2.30 } and { 2.31 }. Thus, numerical values for β_1 and β_2 coefficients that are quite large in absolute value are almost inevitable whenever $|\rho_{12}|$ is near one.

Figure 2.3 Signs of Coefficients in Two Predictor Regression Models when $r_{y2} = 0.5$.



Another perspective on the above observations is provided by the decomposition of β_1 and β_2 into uncorrelated γ components. For example, in the special case we are considering where $\sigma_y = \sigma_1 = \sigma_2$, principal axes again fall at 45° angles with the given predictor axes. As a result, $\beta_1 = (\gamma_1 + \gamma_2) / \sqrt{2}$ and $\beta_2 = (\gamma_1 - \gamma_2) / \sqrt{2}$, where

$$\gamma_1 = (\rho_{y1} + \rho_{y2}) / [(1 + \rho_{12}) \cdot \sqrt{2}]$$

and

$$\gamma_2 = (\rho_{y1} - \rho_{y2}) / [(1 - \rho_{12}) \cdot \sqrt{2}] .$$

Now, note that only γ_2 is sensitive to the numerical difference between ρ_{y1} and ρ_{y2} . Furthermore, only γ_2 increases in absolute size as ρ_{12} approaches +1 ; in fact, γ_1 decreases in absolute value as ρ_{12} increases! And “wrong" sign problems emerge whenever $|\gamma_2|$ exceeds γ_1 .

2.10.2 Actual Practice ...Inferences Conditional on Given Predictor Variables

Now that details of the statistical theory of two-predictor regressions have been outlined, we can explore their implications in real-life applications. In actual practice, the observed numerical value of $r_{12} = \widehat{\rho}_{12}$ (the sample estimate of ρ_{12}) may be more of an artifact of how the observed data were collected than of any sort of measure of the "natural" joint-variation of X_1 and X_2 .

> When our data come from a "designed" experiment, we have hopefully observed responses at a set of pairs of numerical values for X_1 and X_2 with the highly desirable property that $\widehat{\rho}_{12}$ is close to zero!
>
> When our data collection is merely "observational" or we have simply compiled historical (retrospective) results, the implied $\widehat{\rho}_{12}$ could be quite large. This does not imply, however, that X_1 and X_2 "should" or naturally "would" track each other ...with both tending either to increase together or to decrease together from observation-to-observation of the process under study. In other words, we have only observed responses corresponding to a strict subset of the (X_1 , X_2) pairings that would have been explored in any sort of "well-designed" experimental situation.

Conditional upon the given X_1 and X_2 predictor coordinates, $\widehat{\rho}_{12}$ is simply a known constant. But $\widehat{\rho}_{y1}$ and $\widehat{\rho}_{y2}$ remain stochastic given X_1 and X_2 . In fact, their ratio $\Re = \rho_{y1} / \rho_{y2}$ is not only stochastic given X_1 and X_2 but also numerically unstable (due to ill-conditioning) when $\widehat{\rho}_{12}$ is anywhere near to $+1$. In other words, very small numerical changes in the N observed response y-values can result in major changes in the relative-magnitudes and numerical-signs of β_1 and β_2 of equations { 2.30 } and { 2.31 }.

2.11 Tests of General Linear Hypotheses

A "general linear hypothesis" will be denoted by

$$H: \ \mathbf{A}\,\boldsymbol{\beta} = \boldsymbol{\rho} \qquad \{ 2.35 \}$$

where **A** is a known (r x P) matrix with $1 \leq \text{rank}(\mathbf{A}) = r \leq P$ and $\boldsymbol{\rho}$ is a known (r x 1) vector. When r > 1, assume also that $\mathbf{A}\,\boldsymbol{\beta} = \boldsymbol{\rho}$ are self-consistent linear equations.

Note that the equation, $\mathbf{A}\,\boldsymbol{\beta} = \boldsymbol{\rho}$, of { 2.32 } places a restriction on potential values for the $\boldsymbol{\beta}$ vector. Regression practitioners, using estimates of $\boldsymbol{\beta}$ derived from their data, can make statistical inferences about whether this restriction seems plausible. For example, when the conditional distribution of **y** given **X** is assumed to be multivariate normal, the least squares confidence region for $\mathbf{A}\,\boldsymbol{\beta}$ has the following well-known distribution and properties. Let $F(r, n-p-1; \alpha)$ denote the upper $100(1-\alpha)\%$ point of Snedecor's F-distribution, and let s^2 denote the residual mean square for error of equation { 2.22 }. Then the set of all possible vectors of the form **A b** such that

$$(\mathbf{b} - \mathbf{b}^{o})^{T} \mathbf{A}^{T} [\mathbf{A} (\mathbf{X}^{T}\mathbf{X})^{+} \mathbf{A}^{T}]^{+} \mathbf{A} (\mathbf{b} - \mathbf{b}^{o}) / (r\, s^{2}) \;\leq\; F(r, n - p - 1; \alpha) \qquad \{2.36\}$$

is an unbiased confidence region for $\mathbf{A}\boldsymbol{\beta}$ which covers the unknown true value of $\mathbf{A}\boldsymbol{\beta}$ with probability $(1 - \alpha)$ for every $\boldsymbol{\beta}$ and for every σ^2 under normal distribution theory. This region constitutes the surface and interior of an r-dimensional hyperellipsoid "centered" at $\mathbf{A}\,\mathbf{b}^{o}$, the unbiased estimate of $\mathbf{A}\boldsymbol{\beta}$.

The general solution for $\boldsymbol{\beta}$ to a set of self-consistent linear equations $A\,\beta = \rho$ can be written as

$$\boldsymbol{\beta} = \mathbf{A}^{-}\boldsymbol{\rho} + (\mathbf{I} - \mathbf{A}^{-}\mathbf{A})\,\mathbf{z} , \qquad \{2.37\}$$

where $\mathbf{A}^{-}$ is any generalized inverse for the $\mathbf{A}$ matrix and $\mathbf{z}$ is any (P x 1) vector, Rao(1973), pp. 24-25. Thus { 2.34 } will include $\boldsymbol{\beta}$ vectors outside as well as any inside the hyperellipsoid of { 2.33 }. The restricted least squares estimator of $\boldsymbol{\beta}$ under the hypothesis H : $\mathbf{A}\boldsymbol{\beta} = \boldsymbol{\rho}$ of { 2.32 } is the $\mathbf{b}$ vector (call it $\mathbf{b}^{H}$, say) that minimizes the Lagrange multiplier equation

$$\psi (\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\,\mathbf{b})^{T} (\mathbf{y} - \mathbf{X}\,\mathbf{b}) - 2\,\boldsymbol{\eta}^{T} (\mathbf{A}\boldsymbol{\beta} - \boldsymbol{\rho}) . \qquad \{ 2.38 \}$$

In other words, $\partial\,\psi / \partial\,\mathbf{b} = \mathbf{0}$ and $\partial\,\psi / \partial\,\boldsymbol{\eta} = \mathbf{0}$ together imply that

$$\begin{aligned} \mathbf{b}^{H} &= \mathbf{A}^{*}\boldsymbol{\rho} + (\mathbf{I} - \mathbf{A}^{*}\,\mathbf{A})\,\mathbf{b}^{o} , \\ &= \mathbf{b}^{o} - \mathbf{A}^{*} (\mathbf{A}\,\mathbf{b}^{o} - \boldsymbol{\rho}) , \end{aligned} \qquad \{2.39\}$$

where $\mathbf{A}^{*} = (\mathbf{X}^{T}\mathbf{X})^{+} \mathbf{A}^{T} [\mathbf{A} (\mathbf{X}^{T}\mathbf{X})^{+} \mathbf{A}^{T}]^{+}$ is one specific choice for the generalized inverse, $\mathbf{A}^{-}$, in $\mathbf{A}\,\mathbf{A}^{-}\,\mathbf{A} = \mathbf{A}$.

The resulting F-statistic for the test if the hypothesis H : $\mathbf{A}\boldsymbol{\beta} = \boldsymbol{\rho}$ of {2.32} is thus

$$\begin{aligned} F &= (\mathbf{b}^{o} - \mathbf{b}^{H})^{T}\, \mathbf{X}^{T}\mathbf{X} (\mathbf{b}^{o} - \mathbf{b}^{H}) / (r\, s^{2}) , \\ &= (\mathbf{A}\,\mathbf{b}^{o} - \boldsymbol{\rho})^{T}\, [\mathbf{A} (\mathbf{X}^{T}\mathbf{X})^{+} \mathbf{A}^{T}]^{+} (\mathbf{A}\,\mathbf{b}^{o} - \boldsymbol{\rho}) / (r\, s^{2}) \end{aligned} \qquad \{2.40\}$$

with numerator degrees-of-freedom=r and denominator degrees-of-freedom=(N – R – 1).

2.12 Weighted Residual Analyses

An extremely important phase of multiple regression modeling is that in which a practitioner examines fitted residuals to...

(i) uncover evidence of systematic lack-of-fit in the expectation model;

Examples of these sorts of problems include unrecognized “curvature" that might be removed by a well-chosen transformation of response and/or predictor variables and also failure to include certain available explanatory variables.

(ii) identify individual observations exerting undue “influence" on the fit;

Examples here include high “leverage" predictor variable combinations and/or “outlying" (maverick) response values.

and/or to

(iii) detect violations of the assumed variance-covariance structure.

Mild deviations from an assumed homoscedastic, uncorrelated error structure usually may not have any major effects. But blatant deviations from an assumed dispersion structure can make the “usual formulas" quite poor indicators of reality.

Our treatment here will avoid almost all practical aspects of residual analyses; highly readable information on basic strategy/tactics is provided by, say, Chapter 3 of Draper and Smith(1981) or Chapters 5 and 6 of Weisberg(1980) or the book of Belsley, Kuh and Welsch(1980). In fact, all we really want to do here is to display/discuss some fundamental residual analysis formulas that are slightly more general than those available elsewhere:

Here in Section §2.12, we consider a residual analysis formulation suitable for the “weighted" least-squares fits commonly used in “robust" regression and potentially useful in visual re-regression (VRR.) Of course, our weighted formulation does include ordinary (unweighted) least-squares residuals as a special case.

In Section §3.5 at the end of Chapter 3, we discusses analyses of residuals resulting from shrinkage-regression fits. There we show how shrinkage introduces bias into residuals when the expectation model is correct, just as shrinkage introduces bias into coefficient estimates. In fact, shrinkage usually changes the variance, the leverage, and the overall influence of each and every observation!

We start by considering a heteroscedastic variance formulation for multiple regression models that generalizes the homoscedastic variance case described in equation { 2.2 }. The fundamental mechanism involved in this form of “robust" fitting is to reduce the weights assigned to observations that are candidates for outliers in the sense that their fitted residuals are relatively large. We will not consider details of iterative methods for defining these weights here in Chapter 2; these topics are treated in Chapter 9. Instead, we proceed here as if observation weights are given values.

While regression weights are usually viewed as being inversely proportional to the variances of their observations, it certainly is not mandatory to visualize outliers as having high variability. Indeed, outliers can also result because their expected values deviate wildly from the general

pattern of other observations. In any case, assigning a weight of zero to an observation certainly does not imply that that observation has infinite variance. In fact, as detailed below, a weight of zero really implies simply that the expected value and the variance of that observation could be any pair of finite values. On the other hand, assigning strictly positive weights to a subset of observations will be interpreted here as an attempt to make both of the resulting model equations (expected value and variance) fit relatively well to all data points of that subset.

The (un-centered) multiple regression model considered here in Section §2.11 will be:

Conditional Expectation of $\mathbf{u}$ given $\mathbf{Z}$: $\quad E(\mathbf{u} \mid \mathbf{Z}) = \mathbf{Z}\boldsymbol{\beta}$, and $\qquad \{ 2.41 \}$

Conditional Variance of $\mathbf{u}$ given $\mathbf{Z}$: $\quad V(\mathbf{u} \mid \mathbf{Z}) = \sigma^2 \mathbf{W}^{-}$, $\qquad \{ 2.42 \}$

where $\mathbf{Z}$ is $(P+1) \times N$ [with the $\mathbf{1}$ vector as its first column] and $\mathbf{W}^{-}$ is any diagonal, generalized-inverse matrix [Rao(1973), page 24] for the $N \times N$ diagonal and non-negative definite matrix of finite weights, $\mathbf{W}$. I.E. $\mathbf{W}^{-}$ is any diagonal matrix such that $\mathbf{W}\,\mathbf{W}^{-}\mathbf{W} = \mathbf{W}$. In other words, w_{ii}^{-} must equal $1/w_{ii}$ whenever $w_{ii} > 0$, but w_{ii}^{-} can be any finite value whenever $w_{ii} = 0$ for $1 \le i \le N$.

Consider now the transformations

$$\mathbf{y}^* = \mathbf{W}^{1/2}\,\mathbf{u} \quad \text{and} \quad \mathbf{X}^* = \mathbf{W}^{1/2}\,\mathbf{Z}, \qquad \{ 2.43 \}$$

where the diagonal matrix of positive-square-roots, $\mathbf{W}^{1/2}$, is uniquely determined from $\mathbf{W}$. The rows of $\mathbf{y}^*$ and $\mathbf{X}^*$ corresponding to zero weights are thus identically zero; after all, we have explicitly excluded all cases where the corresponding rows of $\mathbf{u}$ and $\mathbf{Z}$ might contain infinite values ($\pm\infty$.)

Our generalized regression models of equations { 2.38 } and { 2.39 } thus imply the models $E(\mathbf{y}^* \mid \mathbf{X}^*) = \mathbf{X}^*\boldsymbol{\beta}$ and $V(\mathbf{y}^* \mid \mathbf{X}^*) = \sigma^2 \cdot \mathbf{D}$, where $\mathbf{D}$ is a $N \times N$ diagonal matrix each of whose diagonal elements is either a zero or a one. [Note that the $\mathbf{D}$ matrix is its own uniquely-determined Moore-Penrose inverse, Rao(1973), page 26.] Letting $N^* \le N$ denote the rank of $\mathbf{D}$, the degrees-of-freedom-for-error in our generalized model are $N^* - P - 1$, at least when $\mathbf{X}^*$ is of full rank, namely $P + 1$. The corresponding least-squares estimates of β and σ^2 from the regression of $\mathbf{y}^*$ onto $\mathbf{X}^*$ are then given by

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{*T}\mathbf{X}^*)^{-1}\,\mathbf{X}^{*T}\mathbf{y}^* = (\mathbf{Z}^T\mathbf{W}\,\mathbf{Z})^{-1}\,\mathbf{Z}^T\mathbf{W}\,\mathbf{u} \qquad \{ 2.44 \}$$

and

$$s^2 = (\mathbf{u} - \mathbf{Z}\widehat{\boldsymbol{\beta}})^T\mathbf{W}(\mathbf{u} - \mathbf{Z}\widehat{\boldsymbol{\beta}}) / (N^* - P - 1). \qquad \{ 2.45 \}$$

Note that $\widehat{\beta}$ is unbiased, $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, and its variance is $V(\widehat{\boldsymbol{\beta}}) = \sigma^2 \cdot (\mathbf{Z}^T\mathbf{W}\,\mathbf{Z})^{-1}$, where $\sigma^2 = E(s^2)$ of { 2.42 }.

The vector of generalized residuals from model equations { 2.38 } and { 2.39 } resulting from the β and σ^2 estimates of equations { 2.41 } and { 2.42 } are, by definition, of the form

$$\widehat{\mathbf{r}}_{\mathbf{u}} \equiv \mathbf{u} - \mathbf{Z}\widehat{\boldsymbol{\beta}} = [\,\mathbf{I} - \mathbf{Q}\,\mathbf{W}\,]\,\mathbf{u} \qquad \{\ 2.46\ \}$$

where

$$\mathbf{Q} = \mathbf{Z}\,(\,\mathbf{Z}^{\mathrm{T}}\,\mathbf{W}\,\mathbf{Z}\,)^{-1}\,\mathbf{Z}^{\mathrm{T}}\,. \qquad \{\ 2.47\ \}$$

The conditional variance-covariance matrix of $\widehat{\mathbf{r}}_{\mathbf{u}}$ given $\mathbf{Z}$ is thus of the form

$$\mathrm{V}(\widehat{\mathbf{r}}_{\mathbf{u}}\,|\,\mathbf{Z}) = \sigma^2 \cdot (\,\mathbf{W}^{-} - \mathbf{Q}\,\mathbf{D} - \mathbf{D}\,\mathbf{Q} + \mathbf{Q}\,)\,. \qquad \{\ 2.48\ \}$$

Note that two very different sorts of formulas for the ii-th diagonal element of this conditional variance matrix result when the weight given to the i-th observation is, respectively, positive or zero:

$$\mathrm{V}(\widehat{\mathrm{r}}_{\mathrm{u(i)}}\,|\,\mathrm{Z}) = \sigma^2 \cdot (\,\mathrm{w}_{\mathrm{ii}}^{-1} - \mathrm{q}_{\mathrm{ii}}\,) \quad \text{when } \mathrm{w}_{\mathrm{ii}} > 0\,, \qquad \{\ 2.49\ \}$$

but

$$\mathrm{V}(\widehat{\mathrm{r}}_{\mathrm{u(i)}}\,|\,\mathrm{Z}) = \sigma^2 \cdot (\,\mathrm{w}_{\mathrm{ii}}^{-} + \mathrm{q}_{\mathrm{ii}}\,) \quad \text{when } \mathrm{w}_{\mathrm{ii}} = 0\,. \qquad \{\ 2.50\ \}$$

Similarly, the conditional covariance between the i-th and j-th element of $\widehat{\mathrm{r}}_{\mathrm{u}}$ given Z is

$$\mathrm{Cov}(\widehat{\mathrm{r}}_{\mathrm{u(i)}}\,,\widehat{\mathrm{r}}_{\mathrm{u(j)}}\,|\,\mathrm{Z}) = -\sigma^2 \cdot \mathrm{q}_{\mathrm{ij}} \quad \text{when } \mathrm{w}_{\mathrm{ii}} > 0 \text{ and } \mathrm{w}_{\mathrm{jj}} > 0\,, \qquad \{\ 2.51\ \}$$

where $\mathrm{q}_{\mathrm{ij}} = \mathbf{z}_{\mathrm{i}}^{\mathrm{T}}\,(\,\mathbf{Z}^{\mathrm{T}}\,\mathbf{W}\,\mathbf{Z}\,)^{-1}\,\mathbf{z}_{\mathbf{j}}$ and $\mathbf{z}_{\mathrm{i}}^{\mathrm{T}}$ is the i-th row of $\mathbf{Z}$. Otherwise, this covariance is zero.

The i-th residual is standardized by dividing it by the "usual" estimate of its standard deviation, the square root of the i-th diagonal element of { 2.45 } with σ^2 estimated by s^2 of { 2.42 } :

$$\mathrm{r}^{\mathrm{s}}_{\mathrm{u(i)}} = \frac{\widehat{\mathrm{r}}_{\mathrm{u(i)}}}{\mathrm{s}\cdot\sqrt{\mathrm{w}_{\mathrm{ii}}^{-1} - \mathrm{q}_{\mathrm{ii}}}} \quad \text{when } \mathrm{w}_{\mathrm{ii}} > 0\,, \qquad \{\ 2.52\ \}$$

but

$$\mathrm{r}^{\mathrm{s}}_{\mathrm{u(i)}} = \frac{\widehat{\mathrm{r}}_{\mathrm{u(i)}}}{\mathrm{s}\cdot\sqrt{\mathrm{w}_{\mathrm{ii}}^{-} + \mathrm{q}_{\mathrm{ii}}}} \quad \text{when } \mathrm{w}_{\mathrm{ii}} = 0\,. \qquad \{\ 2.53\ \}$$

Unfortunately, the numerator and denominator of this ratio are usually not independent statistics when $\mathrm{w}_{\mathrm{ii}} > 0$; thus, standardized residuals corresponding to strictly positive weights generally do not follow Student's-t distribution under normal theory.

The i-th residual is studentized by dividing it by an independent estimate of its standard deviation. The arguments used by Beckman and Trussell(1974) and also by Ellenberg(1973,1976) are easily extended to our uncorrelated-observations, heterogeneous-variances model, equations { 2.38 } and { 2.39 }. Namely, consider the estimate of β resulting from "leaving out" the i-th observation from the model:

$$\widehat{\boldsymbol{\beta}}_{(-i)} \;=\; (\mathbf{Z}^T \mathbf{W} \mathbf{Z} - w_{ii} \cdot \mathbf{z}_i \mathbf{z}_i^T)^{-1} (\mathbf{Z}^T \mathbf{W} \mathbf{u} - w_{ii} \cdot z_i \cdot u_i) , \qquad \{ 2.54 \}$$

where $\mathbf{z}_i^T$ is the i-th row of $\mathbf{Z}$. Note that

$$\widehat{\beta} - \widehat{\beta}_{(-i)} \;=\; (\mathbf{Z}^T \mathbf{W} \mathbf{Z})^{-1} z_i \cdot \frac{w_{ii} \cdot (u_i - \mathbf{z}_i^T \widehat{\beta})}{[1 - w_{ii} \cdot q_{ii}]} . \qquad \{ 2.55 \}$$

Thus $w_{ii} = 0$ in { 2.51 } and { 2.52 } immediately implies that $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{(-i)}$. When $w_{ii} > 0$, the residual mean square, $s^2_{(-i)}$, resulting from leaving-out the i-th observation still yields an unbiased estimator of the error variance, σ^2. More importantly, this leave-out-the-i-th-observation estimate will be statistically independent of that i-th observation. Furthermore, a very simple formula for $s^2_{(-i)}$ in terms of s^2 from the full model and the i-th residual from the full model is:

$$(N^* - P - 1 - d_{ii}) \cdot s^2_{(-i)} = (N^* - P - 1) s^2 \; - \widehat{r}^2_{u(i)} / (1 - w_{ii} \cdot q_{ii}) . \qquad \{ 2.56 \}$$

In particular, $w_{ii} = 0$ implies that $d_{ii} = 0$, that the i-th residual from the regression of $\mathbf{y}^*$ onto $\mathbf{X}^*$ is zero. But $w_{ii} > 0$ implies that $d_{ii} = 1$ and that $s^2_{(-i)} \neq s^2$. The studentized residuals are thus of the general form:

$$\widehat{t}_{u(i)} \;=\; \frac{\widehat{r}_{u(i)}}{s_{(-i)} \cdot \sqrt{w_{ii}^{-1} - q_{ii}}} \;=\; r^s_{u(i)} \cdot \sqrt{\frac{N^* - P - 2}{N^* - P - 1 - [r^s_{u(i)}]^2}} \quad \text{when } w_{ii} > 0 ,$$

$$\{ 2.57 \}$$

but

$$\widehat{t}_{u(i)} \;=\; \frac{\widehat{r}_{u(i)}}{s \cdot \sqrt{w_{ii}^{-} + q_{ii}}} \;=\; r^s_{u(i)} \;\text{ when } w_{ii} = 0 . \qquad \{ 2.58 \}$$

The Cook(1977) measure of the overall influence of the i-th observation upon the regression is thus

$$\text{CINFL}_i \qquad = [\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(-i)}]^T \mathbf{Z}^T \mathbf{W} \mathbf{Z} [\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(-i)}] / (P + 1) \cdot s^2 \qquad \{ 2.59 \}$$

$$= \frac{w_{ii}^2 \cdot (u_i - \mathbf{z}_i^T \widehat{\boldsymbol{\beta}})^2 \cdot [\mathbf{z}_i^T (\mathbf{Z}^T \mathbf{W} \mathbf{Z})^{-1} \mathbf{z}_i]}{[P+1] s^2 \cdot [1 - w_{ii} \cdot \mathbf{z}_i^T (\mathbf{Z}^T \mathbf{W} \mathbf{Z})^{-1} \mathbf{z}_i]^2} .$$

Note, in particular, this influence is $\text{CINFL}_i = 0$ when $w_{ii} = 0$.

Now, defining the leverage of the i-th observation on the regression to be $\Lambda_i = 0$ when $w_{ii} = 0$ and, otherwise, to be

$$\Lambda_i = \frac{\text{Predictive Variance}}{\text{Residual Variance}} = \frac{\mathbf{z}_i^T (\mathbf{Z}^T \mathbf{W} \mathbf{Z})^{-1} \mathbf{z}_i}{[\, w_{ii}^{-1} - \mathbf{z}_i^T (\mathbf{Z}^T \mathbf{W} \mathbf{Z})^{-1} \mathbf{z}_i]} \qquad \{ 2.60 \}$$

we can rewrite Cook's measure of influence as

$$\begin{aligned} \text{CINFL}_i &= 0 && \text{when } w_{ii} = 0, \\ &= \frac{(\, r^s_{u(i)})^2 \cdot \Lambda_i}{[\, P+1\,]} && \text{otherwise,} \end{aligned} \qquad \{ 2.61 \}$$

which is proportional to the product of the leverage times the square of the standardized residual.

The results derived here in Section §2.12 on analysis of weighted residuals can be summarized as follows:

When the weight given to an observation is zero, the corresponding residual can be visualized as having arbitrary variance, $V(\hat{\tau}_{u(i)} \,|\, \mathbf{Z}) = \sigma^2 \cdot (\, w_{ii}^{-} + q_{ii}\,)$ for any generalized inverse of $w_{ii} = 0$. That observation then has zero influence, zero leverage, and its studentized and standardized residuals coincide.

When the weight given to an observation is positive, the corresponding residual has variance $V(\hat{\tau}_{u(i)} \,|\, \mathbf{Z}) = \sigma^2 \cdot (\, w_{ii}^{-1} - q_{ii}\,)$ as in $\{ 2.45 \}$ and $\{ 2.46 \}$. Such an observation will have positive measures of influence and of leverage, and the corresponding studentized and standardized residuals will usually not coincide.

References for Chapter 2

Beckman, R. J. and Trussell, H. J. (1974). "The distribution of an arbitrary studentized residual and the effects of updating in multiple regression." **Journal of the American Statistical Association** 69, 199-201.

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). **Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.** New York: John Wiley.

Cook, R. D. (1977). "Detection of influential observations in linear regression." **Technometrics** 19, 15-18.

Draper, N. R. and Smith, H. (1981). **Applied Regression Analysis**, Second Edition. New York: John Wiley.

Ellenberg, J. H. (1973). "The joint distribution of the standardized least squares residuals from a general linear regression." **Journal of the American Statistical Association** 68, 941-943.

Ellenberg, J. H. (1976). "Testing for a single outlier from a general linear regression." **Biometrics** 32, 637-645.

Massy, W. F. (1965). "Principal components regression in exploratory statistical research." **Journal American Statistical Association** 60, 234-256.

Obenchain, R. L. (1975a). "Residual optimality: ordinary vs. weighted vs. biased least squares." **Journal of the American Statistical Association** 70, 407-416.

Obenchain, R. L. (1975b). "Ridge analysis following a preliminary test of the shrunken hypothesis." **Technometrics**, 17, 431-441. (Discussion: McDonald, G. C., 443-445.)

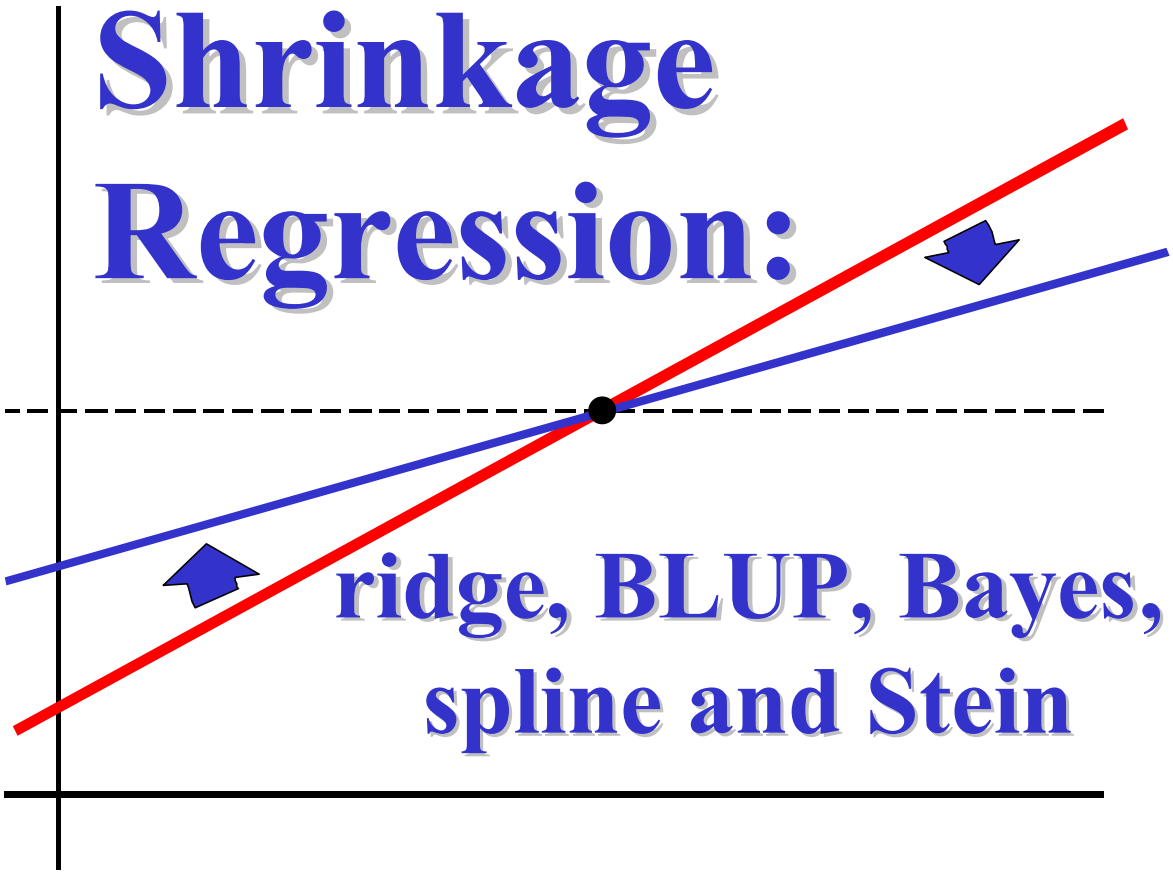
Obenchain, R. L. (1976). "Methods of ridge regression." **Proceedings of the Ninth International Biometric Conference**, Invited Papers, Volume One, 37-57, Boston.

Obenchain, R. (1977). "Classical F-tests and confidence regions for ridge regression." **Technometrics** 19, 429-439.

Rao, C. R. (1973). **Linear Statistical Inference and its Applications, 2nd edition.** New York: John Wiley & Sons.

Smith, G. and Campbell, F. (1980). "A critique of some ridge regression methods" (with discussion.) **Journal of the American Statistical Association** 75, 74-103.

Weisberg, S. (1980). **Applied Linear Regression.** New York: John Wiley.



Chapter 03: Shrinkage Regression Fundamentals

Bob Obenchain, Ph.D.
softR$_X$ freeware
13212 Griffin Run
Carmel, Indiana 46033-8835

Chapter 3: SHRINKAGE REGRESSION FUNDAMENTALS

The regression estimators of main interest to our exposition here are known as generalized shrinkage estimators or generalized ridge regression estimators. The vector of estimators for all P of the elements of the β coefficient vector in a linear model, such as that in equations { 2.1 } and { 2.2 } of Chapter 2, will be denoted here by the symbol $b^{\bigstar}$ and will be of the general form

$$b^{\bigstar} \;=\; G\,\Delta\,c \;\;=\; \sum_{j=1}^{j=R} \vec{g}_j \cdot \delta_j \cdot c_j \;. \qquad \{\, 3.1 \,\}$$

In equation { 3.1 }, $\vec{g}_j$ is the j-th column of the principal-axis direction-cosines matrix, G ; δ_j is the j-th diagonal element of the shrinkage factors matrix, Δ ; c_j is the j-th element of the uncorrelated components vector, c , of equation { 2.16 } ; and R is the rank of the centered regressor X matrix. We will usually restrict the RANGE of interest for all R of the shrinkage factors, $\delta_1 \,, \ldots, \delta_R$, to...

$$0 \;\leq\; \delta_j \;\leq\; 1 \;. \qquad \{\, 3.2 \,\}$$

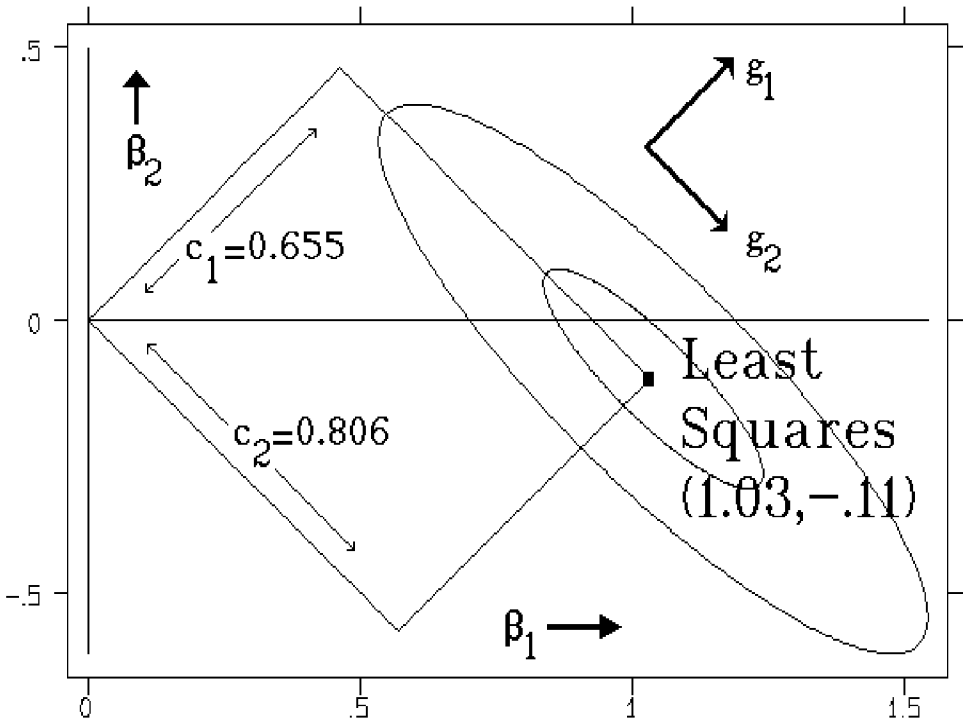
Kronecker's delta function, which will be familiar to many readers, takes on only two possible values, zero and one. In our exposition, shrinkage δ factors can take on all numerical values between zero and one, inclusive. But our use of this δ notation may help readers remember the conventional extreme values, 0 and 1, of the range allowed for shrinkage factors.

The generalized shrinkage estimator corresponding to $\delta_1 = \ldots = \delta_R = 1$ (i.e. $\Delta = I$) coincides with the ordinary least squares estimator, $b^{\bigstar} = b^{o}$. And the shrinkage estimator corresponding to $\delta_1 = \ldots = \delta_R = 0$ (i.e. $\Delta = 0$) is $b^{\bigstar} = \vec{0}$. When there is no restriction on shrinkage factors other than $0 \leq \delta_j \leq 1$ for $1 \leq j \leq R$, the set of all possible generalized shrinkage estimators, $b^{\bigstar}$, constitutes the surface and interior of an R-dimensional hyper-rectangle with...

– each of its edges parallel to a corresponding principal axis of the given regressors,
– one of its 2^R vertices at the least squares estimate, b^o, and
– the "diagonally opposite" vertex at the shrinkage origin, which is usually $\vec{0}$.

The above generalized shrinkage "geometry" is illustrated below, in Figure 3.1, for the P=R=2 dimensional case described in the simple N = 10 observation numerical example discussed in Sections §2.3, §2.6 and §2.10 of Chapter 2.

Figure 3.1 Two-Dimensional Generalized Shrinkage Rectangle



Again, all points either on the boundary or in the interior of the rectangle of Figure 3.1 are generalized shrinkage estimators as defined by equation { 3.1 }.

3.1 Moments of Generalized Shrinkage Estimators

Generalized shrinkage estimators, { 3.1 }, can be linear estimators of β for the fixed-effect model of equation { 2.1 } (or { 2.3 }) of Chapter 2. But they are linear only in cases where all

R of the generalized shrinkage factors, $\delta_1, \ldots, \delta_R$, are non-stochastic given X. In these special cases, the conditional expected value of $b^{\bigstar}$ will be

$$E(b^{\bigstar} \mid X) = G \Delta \gamma . \qquad \{ 3.3 \}$$

These shrinkage estimators, $b^{\bigstar}$, are generally "biased" estimators. After all, the shrinkage expectation vector is $G \Delta \gamma$, and this vector is generally $\neq \beta = G \gamma$ in cases where $\Delta \neq I$. The bias in $b^{\bigstar}$, namely $G (I - \Delta) \gamma$, is usually unknown in practical applications because, just like β itself, the true $\gamma = G^T \beta$ components would also be unknown.

Similarly, the conditional variance matrix of $b^{\bigstar}$ for non-stochastic shrinkage in a fixed-effect model is

$$V(b^{\bigstar} \mid X) = \sigma^2 \cdot G \Delta^2 \Lambda^{-1} G^T . \qquad \{ 3.4 \}$$

Expression { 3.4 } is certainly not a universally valid variance formula; it simply does not apply when the shrinkage factor matrix, Δ, is actually stochastic given X. And uncertainty about an appropriate form and extent for shrinkage occurs in most (if not all) practical applications! After all, a shrinkage practitioner generally chooses his/her desired shrinkage "pattern" only after examining several functions (statistics) that depend upon the observed response vector, y. For example, the practitioner may make his/her choice only after visually examining shrinkage TRACE displays ...looking for shrinkage that will "stabilize" coefficients and/or change the numerical signs of fitted coefficient estimates! Exact moment formulae for the resulting NONLINEAR shrinkage estimators frequently cannot be computed. In fact, it is usually difficult to conjecture whether equation {3.4} is even approximately "correct" in situations where one's choice of Δ shrinkage factors ends up being stochastic. However, nothing less than the above sorts of visualization strategies/tactics give sufficient insights to make "realistic" shrinkage decisions in almost all practical shrinkage-regression applications.

3.2 Shrinkage Inflation of the Residual Mean Square

The squared length of the residual vector corresponding to a shrinkage estimator $b^{\bigstar} = G \Delta \gamma$ is used to define the corresponding "inflated" residual mean square, $RMS^{\bigstar}$, as follows...

$$\begin{aligned} RMS^{\bigstar} &= [\,(y - X b^{\bigstar})^T (y - X b^{\bigstar})\,] / (N - R - 1), \qquad \{ 3.5 \} \\ &= [\; y^T (I - H \Delta H^T)^2 y \;] / (N - R - 1), \\ &= [\; y^T (I - H H^T) y + y^T H (I - \Delta)^2 H^T y \;] / (N - R - 1), \\ &= RMS^{o} + [\, y^T y / (N - R - 1) \,] \cdot r^T (I - \Delta)^2 r , \end{aligned}$$

where $RMS^{o} = s^2 = y^T (I - H H^T) y / (N - R - 1)$ is the least-squares residual mean square for error. Note that the $(N - R - 1)$ factor in the denominator of { 3.5 } makes RMS^{o} an unbiased estimator of σ^2 in equations { 2.2 } and { 2.4 } under normal distribution theory, Johnson and Kotz(1970), equation (10), page 168. Furthermore, by the "Principle of Least Squares," RMS^{o} is (essentially by its very definition) the minimum residual-mean-square. In

fact, the shrinkage residual-mean-square, RMS$^{\bigstar}$ of { 3.5 }, is usually a fairly "uninteresting" statistic that may grossly overestimate the true σ^2.

3.3 The Hoerl-Kennard "Ordinary" Ridge Family

The ridge estimators originally proposed by Hoerl and Kennard (1970a,b) are of the highly specialized form:

$$b^{\bigstar} \quad = \quad (X^T X + k \cdot I)^{-1} X^T y . \qquad \{ 3.6 \}$$

However, simple matrix-algebraic manipulations, using the singular value decomposition of X in equation { 2.8 }, show that these estimators can be rewritten as...

$$\begin{aligned} b^{\bigstar} \quad &= \quad [G (\Lambda + k \cdot I) G^T]^{-1} G \Lambda^{1/2} H^T y , \\ &= \quad G (\Lambda + k \cdot I)^{-1} \Lambda^{1/2} H^T y , \\ &= \quad G [(\Lambda + k \cdot I)^{-1} \Lambda] \Lambda^{-1/2} H^T y , \\ &= \quad G \Delta c \end{aligned}$$

for $\Delta = (\Lambda + k \cdot I)^{-1} \Lambda$ and $k \geq 0$. Equivalently, the corresponding generalized shrinkage δ-factors are...

$$\delta_j \quad = \quad \lambda_j / (\lambda_j + k) \quad \text{for } 1 \leq j \leq R.$$

The shrinkage family of equation { 3.6 } is easily derived as a solution to either of two optimization problems:

(1) What is the locus of the most-likely β estimate vectors for each given length?

and

(2) What is the locus of shortest β estimate vectors for each given likelihood?

For example, the Lagrange equation with multiplier k for maximizing the likelihood that a vector of values b is β (i.e. minimizing the corresponding residual sum-of-squares) subject to the restriction that the squared length of b is $b^T b = C^2$ can be written as

$$\Psi(b) = (y - Xb)^T (y - Xb) + k \cdot (b^T b - C^2) ,$$

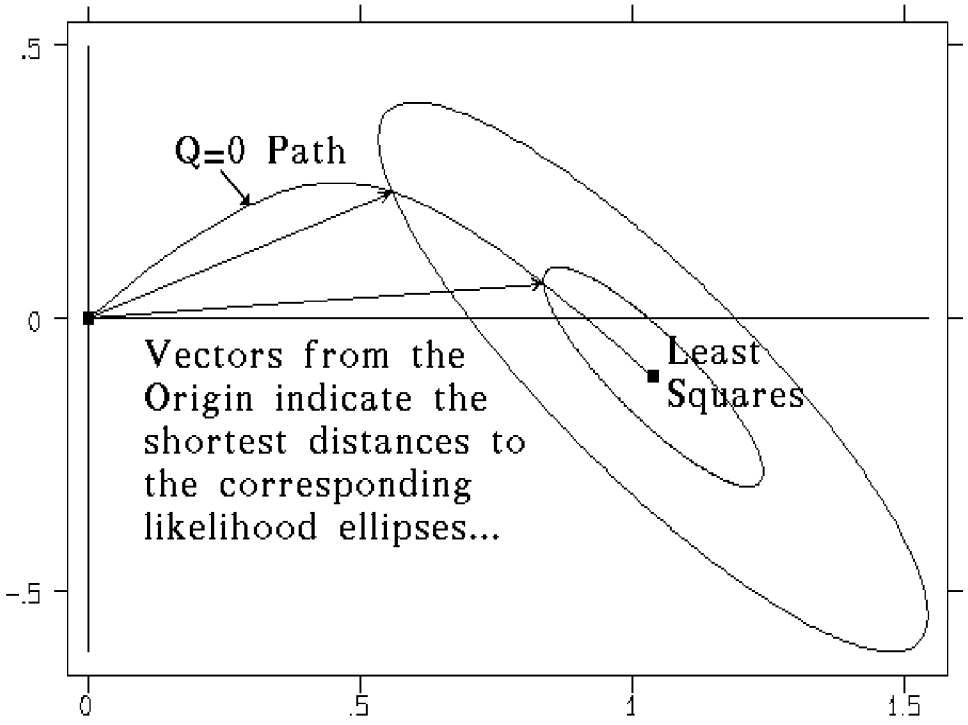
where the X matrix and the y vector have been "centered" as in equations { 2.3 } and { 2.4 }. Equating $\partial \Psi / \partial b = 2 \cdot (X^T X b + k \cdot b - X^T y)$ to $\vec{0}$ then yields equation { 3.6 }. The "analytical geometry" of this situation is illustrated in Figure 3.2, below, for the simple $P = R = 2$ dimensional numerical example of Figure 3.1 and Chapter 2. The Hoerl-Kennard "ordinary" ridge shrinkage path is labeled "Q = 0" in Figure 3.2 for consistency with the notation to be introduced in Section §3.3, below.

Note also that equation { 3.6 } offers direct, intuitive appeal whenever a regression problem is ill-conditioned! Namely, it seems to be almost "obvious" that adding small, positive values (

$k \cdot I$) to the diagonal of $X^T X$ will make any nearly singular regressor inner-products matrix "easier" to invert; see Piegorsch and Casella(1989) for information on this and other early motivations for "ridge" regression terminology.

Equation { 3.6 } might seem to suggest that the ($X^T X + k \cdot I$) matrix be re-inverted each time the numerical value of k is changed. In stark contrast, Equation { 3.1 }, which (as we have already seen) includes { 3.6 } as a special case, offers distinct computational advantages! Having once computed the least squares decomposition

Figure 3.2 The Hoerl-Kennard Shrinkage Path



$b^o = G c$ of { 2.14 }, equation { 3.1 } suggests that shrinkage estimates be calculated by simple matrix and vector multiplications, as $b^\star = G \Delta c$. Meanwhile, the great "intuitive" advantage of equation { 3.1 } is that $b^\star = G \Delta c$ is immediately seen as resulting from different amounts of shrinkage along each principal axis of regressors. Finally, the "ordinary" ridge formula $\delta_j = \lambda_j / (\lambda_j + k) = 1 / (1 + k/\lambda_j)$ is seen to have an additional, intuitive appeal. Namely, for each fixed k value, greater shrinkage (a smaller δ_j factor) is applied to the least squares components corresponding to the smallest "spreads" (smallest λ_j values) in the given regressor coordinates.

3.4 The Two-Parameter Generalized Ridge Family

Unlike the totally unrestricted approach of { 3.1 }, the shrinkage factors $(\delta_1, \ldots, \delta_R)$ for most "families" of interest are functions of at most two parameters. The pair of shrinkage "hyper-parameters" we discuss below (MCAL and QPAR) are...

(i) not only adequate to control both the form (or general shape) and the extent of shrinkage,

(ii) but also general enough to include, as special cases, the shrinkage families considered by the vast majority of authors who have published descriptions of generalized shrinkage strategy/tactics.

First of all, as shrinkage occurs, $b^{\bigstar}$ generally moves away from b^{o} and toward $\vec{0}$. The primary shrinkage parameter can thus be taken to be:

$$\begin{aligned} \text{MCAL} &= \text{M}, \quad \text{the "multicollinearity allowance" parameter that controls the extent of shrinkage,} \\ &= R - \delta_1 - \delta_2 - \ldots - \delta_R. \end{aligned} \qquad \{ 3.7 \}$$

Shrinkage of $b^{\bigstar}$ from b^{o} towards $\vec{0}$ follows a "path" whose general shape can also be controlled by the regression practitioner. Our secondary shrinkage parameter will be denoted by:

$$\text{QPAR} = \text{Q}, \quad \text{the "shape" parameter that controls the curvature of the shrinkage path through regression coefficient likelihood space.}$$

Specifically, the primary 2-parameter functional form for shrinkage factors considered here will be:

$$\begin{aligned} \delta_j &= \lambda_j \; / \; [\; \lambda_j + \text{Konst.} \; \lambda_j^{Q} \;] \\ &= 1 \; / \; [\; 1 + \text{Konst.} \; \lambda_j^{Q-1} \;], \end{aligned} \qquad \{ 3.8 \}$$

where the constant, Konst, in { 3.8 } is chosen to provide any specified extent of shrinkage, as quantified by $M = R - \delta_1 - \delta_2 - \ldots - \delta_R$ of { 3.7 }. One interesting implication of this 2-parameter family is this: Unless $Q = 1$, Konst. $= 0$, or Konst. $= +\infty$, a pair of delta shrinkage factors can be equal, $\delta_i = \delta_j$, if and only if their corresponding regressor eigenvalue (sum-of-squares) are also equal, $\lambda_i = \lambda_j$. Because regressor eigenvalues are "usually" distinct (except, say, for designed experiments), the shrinkage estimators in the 2-parameter family of { 3.8 } "usually" do a different amount of shrinkage along each principal axis of regressors.

The two-parameter family of { 3.8 } was apparently first published in Goldstein and Smith (1974), equation (13), where our Q parameter is $1 - m$ in their notation (and m was explicitly

restricted to integer values.) Equation (14) of Goldstein and Smith(1974) and our { 3.8 } can both be rewritten as:

$$b^{\bigstar} = [\, X^T X + k \cdot (\, X^T X \,)^Q \,]^{-1} X^T y \,. \qquad \{ 3.9 \}$$

Restriction of Q to integer values is not necessary, of course. I, personally, have found a somewhat finer mesh of Q-shapes (including at least half-integer values) useful in practical applications. On the other hand, it would seem rather silly, at least to me, to search for a "best" Q-shape to within, say, three or even more decimal places!

Considerable confusion about the 2-parameter family of { 3.8 } has been voiced in statistical literature. For example, Hoerl and Kennard(1975) point out that this family (where our Q is $-$ q in their notation) was proposed by R. W. Somers in a 1964 presentation at an A.I.Ch.E. Symposium in Memphis, Tennessee. Hoerl and Kennard(1975) suggest the restriction that Q be ≤ 0 and integral, and they also express doubts, without really saying why, that this generalization of their original $Q = 0$ proposal will be of any value in practical applications. Draper and Van Nostrand(1979) only heighten this confusion by failing to recognize the relationship between our equations { 3.8 } and { 3.9 } and their equations (3.10) and (3.29). And they repeat the $Q \leq 0$ restriction. We will demonstrate below [and in our shrinkage regression case studies] that the full 2-parameter family of { 3.8 } is not only extremely versatile but also quite useful in actual practice.

To illustrate generalized shrinkage paths of different Q $-$ shapes in Figure 3.3 below, we again return to the P=R=2 dimensional, $N = 10$ observation numerical example discussed in Chapter 2 and used in Figures 3.1 and 3.2. Besides the Hoerl-Kennard ($Q = 0$) path displayed in Figure 3.2, Figure 3.3 also explicitly shows the paths for the $Q = +2$, $Q = +1$, and $Q = -1$ shapes. Furthermore, the two upper edges of the generalized shrinkage rectangle represent the limiting Q-shape path as Q approaches $-\infty$, while the two lower edges represent the limiting Q-shape as Q approaches $+\infty$.

The following classification of numerical values for the QPAR = Q path shape parameter of { 3.8 } and { 3.9 } is of at least historical interest...

(a) QPAR > 1.0 yields what are commonly called "increasing" δ factors.

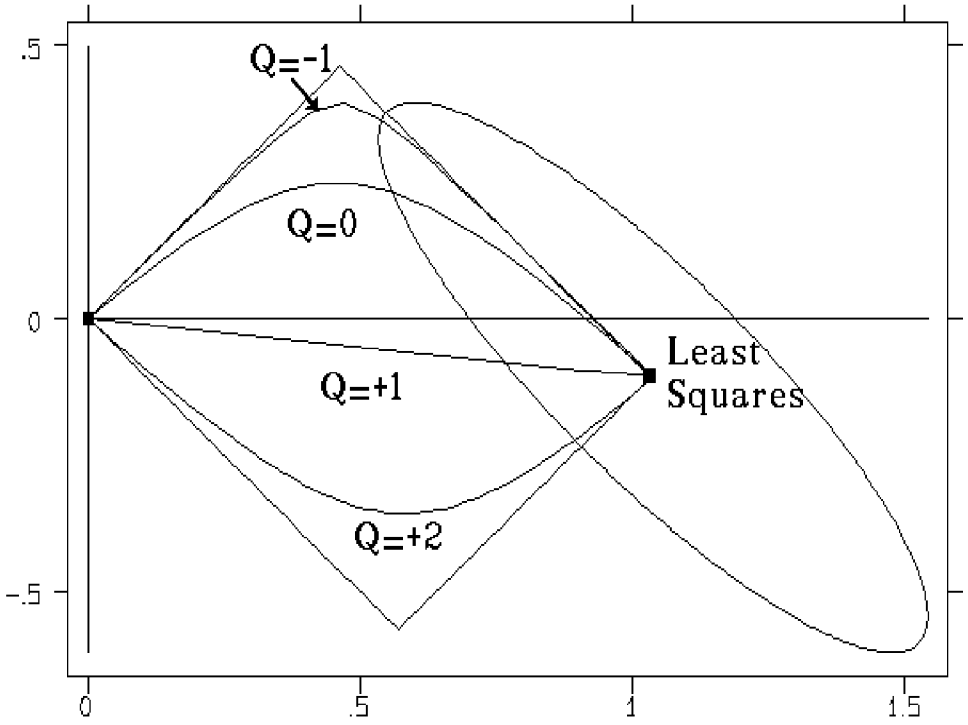
Technically, QPAR > 1.0 implies only that $\delta_1 \leq \ldots \leq \delta_R$. But, again, two δ factors can be equal only when their corresponding eigenvalues are equal. Thus QPAR $>$ 1.0 frequently implies $\delta_1 < \ldots < \delta_R$.

These shrinkage patterns strike me as being somewhat counterintuitive; they actually shrink the relatively precise components of b^o more than its relatively imprecise components. Authors like Thisted(1976), Strawderman(1978) and Casella(1980, 1985) underscore the following paradox: minimax ridge estimators tend to concentrate shrinkage along whichever axes the practitioner regards as LEAST important for improvements in mean-squared-error. Minimax shrinkage patterns with QPAR > 1.0

(shrinkage primarily along major axes) can thus result when the user thinks that he/she has emphasized risk minimization along minor axes!

Strawderman(1978) points out that the two loss functions most commonly used in statistical decision theory correspond to QPAR = 2 and QPAR = 1; for more details, see Section §9.7.

Figure 3.3 Several Q-Shape Shrinkage Patterns



(b) QPAR = 1.0 yields δ factors that are all equal, $\delta_1 = \ldots = \delta_R$.

BLUP estimates for "balanced" designs as well as Stein-type estimates are usually of this special form. Mayer and Willke(1972) called them "shrunken" estimators, but the terminology "uniformly shrunken" estimators would, of course, be more descriptive.

The coefficient trace display is visually uninteresting when the Q – shape is + 1. This trace then consists of P straight lines, each running from a least squares estimate at M = 0 directly to ZERO at M = P. In other words, the relative magnitudes of the elements of $b^{\bigstar}$ are always exactly identical to those of b^{o} in the QPAR = + 1 special case.

Similarly, the trace of shrinkage factors is visually uninteresting when QPAR=1; all P shrinkage factors plot right on top of each other! All shrinkage factors fall on the straight line from $\delta_j = 1$ at M = 0 directly to $\delta_j = 0$ at M = R.

On the other hand, trace plots of estimated mean-squared-error, excess eigenvalues, and the inferior direction can still lead to new insights even when the QPAR shape is $+1$.

(c) QPAR < 1.0 yields what are commonly called "declining δ factors."

Technically, QPAR < 1.0 implies only that $\delta_1 \geq \delta_2 \geq \ldots \geq \delta_R$, but (again) two δ factors can be equal only when the corresponding eigenvalues are equal. Thus QPAR < 1.0 frequently implies $\delta_1 > \ldots > \delta_R$.

QPAR = 0.0 is, of course, the original, "ordinary" form of ridge regression suggested by Hoerl and Kennard(1970a,b); see equation { 3.6 }.

QPAR = -1.0 is an option that yield's delta's which decline more markedly than in the original Hoerl-Kennard formulation; see also Crone(1972) and remarks by Goldstein and Smith (1974) on increasing sensitivity to the "eigenvalue spectrum." Q = -1 yield δ factors of the form...

$$\delta_j = 1 \; / \; [\; 1 + \text{Konst.} \; \lambda_j^{-2} \;] \qquad \{ 3.10 \}$$

(d) The limit as QPAR approaches $-\infty$ (minus infinity) yields "principal-components regression" estimates.

When the ordered eigenvalues of regressor spread are strictly decreasing ($\lambda_1 > \lambda_2 > \ldots > \lambda_R$ without ties), values of QPAR that are negative and large yield shrinkage patterns that are very close, numerically, to what is commonly called Principal-Components Regression; see Kendall(1957) and Massy(1965), method "a", page 241. This "extreme" shrinkage pattern corresponds to moving along a certain chain of edges of the shrinkage-factor hypercube leading from the $\Delta = I$ vertex to the diagonally opposite $\Delta = 0$ vertex. Equivalently, this is the special case where shrinkage factors are non-increasing ($\delta_1 \geq \delta_2 \geq \ldots \geq \delta_R$), and all shrinkage factors, except at most one, are always either 1 or 0 at each point along the path. Marquardt(1970) terms these same estimates "fractional-rank" or "generalized inverse" estimates.

The coefficient TRACE for this shrinkage path consists of broken but connected line segments, with break-points at every integer value of MCAL = M.

For completeness, we note that a 2-parameter family of shrinkage factors that can be quite different from those of { 3.8 } is given by:

$$\delta_j = \min(1, \text{Konst} \cdot \lambda_j^{Q}) \qquad \{ 3.11 \}$$

where the constant, Konst, in { 3.11 } is chosen to provide any specified extent of shrinkage, again quantified by $M = R - \delta_1 - \delta_2 - \ldots - \delta_R$ of { 3.7 }. For a fixed power Q, shrinkage starts in { 3.11 } with any sufficiently large Konst value so that all δ factors will equal 1. This initial Konst value can be taken to be max(λ_1^{-Q}, λ_R^{-Q}). One or more δ factor then starts decreasing as Konst decreases, but at least one δ factor remains fixed at 1 until Konst drops below min(λ_1^{-Q}, λ_R^{-Q}). From this point onward, all δ-factors decrease at the same rate, and all shrunken coefficients have fixed relative magnitudes, converging to zero at Konst = 0. Note the following special path-shapes in this secondary 2-parameter family...

Q = 0.0 for uniform (Stein-like) shrinkage,

Q = 0.5 leads to shrunken coefficients that become both uncorrelated and homoscedastic, as in the equity estimator of Krishnamurthi and Rangaswamy(1987,1989),

and

Q = 1.0 leads to shrunken coefficients that approach the exact same relative magnitudes as the marginal inner-products vector (and are thus guaranteed to have no coefficients with "wrong" signs), as in Obenchain(1978) and equation { 4.17 }.

The secondary 2-parameter family of { 3.11 } strikes me as being somewhat less versatile and somewhat more cumbersome to apply than the primary family of { 3.8 }. For example, closed-form expressions for maximum likelihood estimates within the primary family will be introduced in Chapter 5, and these statistics greatly facilitate choice of Q-shape (as well as shrinkage extent) within the primary family. By way of contrast, no such closed-form expressions are available for the secondary family; choice of Q-shape is thus left either to personal preference or to relatively tedious numerical searches along a variety of different path shapes.

I certainly hope that readers will not be too confused by use of the SAME symbols (Q and Konst) and terminology for BOTH of the 2-parameter families considered here in Section §3.4. These two families are really quite different! For example, uniform shrinkage is Q=1 in the primary family but Q=0 in the secondary family. Similarly, ridge shrinkage requires the Konst to increase in the primary family but to decrease in the secondary family! Enough said?

3.5 The Implicit Intercept for Shrinkage

Most of our attention, so far, has been focused upon the generalized shrinkage estimates, $b^{\bigstar}$, for the P elements of β corresponding to non-constant regressor variables. The resulting "implicit" estimate for the intercept term, μ, is $\bar{y} - \bar{x}^T b^{\bigstar}$ for each $b^{\bigstar}$. (We will see in Chapter 5,

equation { 5.3 }, that this is the Normal-distribution-theory maximum likelihood estimate of μ corresponding to any $b^{\bigstar}$ estimate of β.) Note that this intercept estimate will usually approach $\bar{y}$, NOT zero, as the shrinkage δ-factors approach zero. In other words, shrinkage of regression coefficients to zero can also be visualized as simply a rotation of the fitted regression hyperplane about the $(\bar{x}^T, \bar{y})$ point so that it becomes more horizontal (or "flat") along all P regressor coordinate axes. Generally speaking, the point-of-view adopted here is that the intercept estimate changes in shrinkage regression only because the estimates of the β coefficients are changing. This perspective is illustrated in Figure 3.4 below.

If a formula like { 3.6 } were used to estimate (μ, β^T) without first "centering" either the response y vector or the augmented regressor (1, X) matrix, namely

$$\begin{pmatrix} \mu^{\bigstar} \\ b^{\bigstar} \end{pmatrix} = \left\{ \begin{bmatrix} N & 1^T X \\ X^T 1 & X^T X \end{bmatrix} + k \cdot I \right\}^{-1} \begin{pmatrix} 1^T y \\ X^T y \end{pmatrix},$$

then the $\mu^{\bigstar}$ intercept term would be shrunk to zero just like the $b^{\bigstar}$ coefficient estimates. This sort of situation is illustrated in Figure 3.5, above. Note that shrinkage of this "nonstandard" form can quickly become drastic in the sense that the fit can "miss" all of the data! In other words, all observed data points ultimately end up on the same side of the fitted, shrinkage hyperplane (i.e. on the same side of the fitted line in the $P = R = 1$ case shown in Figure 3.5).

Figure 3.4 The Implicit Shrinkage Intercept

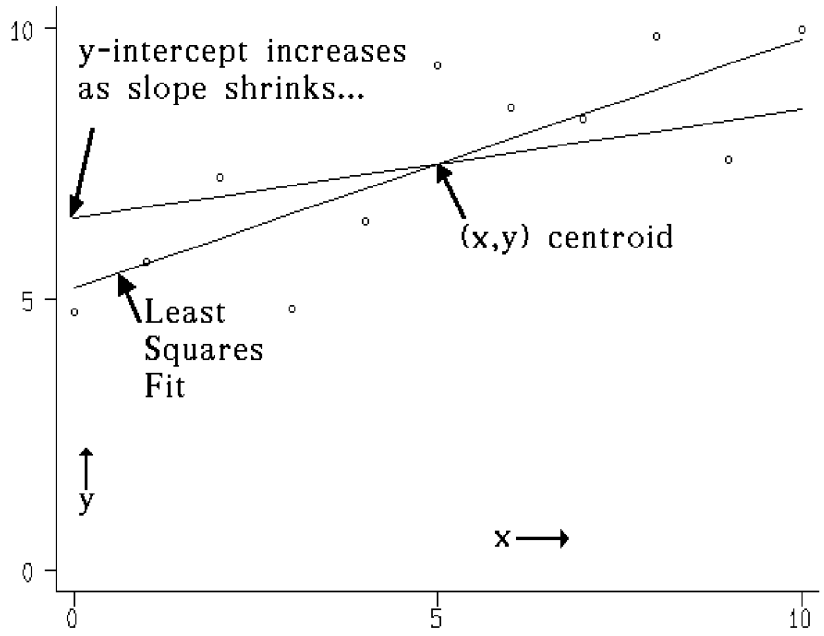
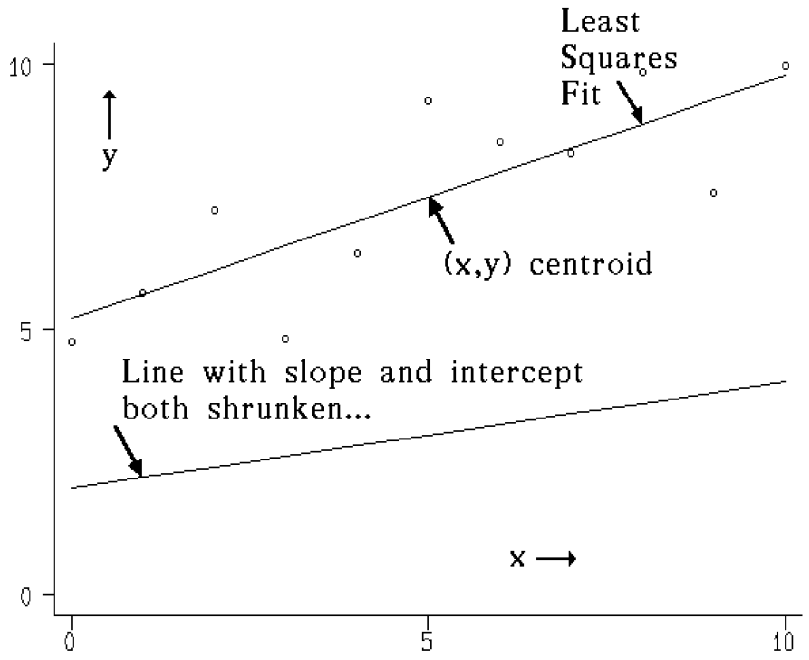


Figure 3.5 Shrinking Both Intercept and Slope



3.6 Models Without an Intercept Term

Models without an intercept, the $1 \cdot \mu$ term of formula { 2.1 }, cannot be analyzed "exactly" using "centered" variables, as in formulas { 2.3 } and { 2.4 }. (Technically, even when the explicit $1 \cdot \mu$ term is absent, the model still actually includes an intercept whenever 1 lies within the column space of the non-constant regressors X matrix BEFORE it has been "centered.") Models without an intercept restrict the regression fit to pass through $y = 0$ at $x = \vec{0}$ instead of through $y = \overline{y}$ at $x = \overline{x}$. In fact, centering can be visualized as a convenient mechanism for assuring that the ($\overline{x}^T$, $\overline{y}$) pivot point is translated so as to coincide with ($\vec{0}^T$, 0). One looses only a single degree-of-freedom for error in estimating β by pre-multiplying both the response y vector and the non-constant regressor X matrix of a model with no intercept on the left by the "centering" projection matrix ($I - 1\,1^T / N$). And we argue below that, in essence, using this wrong (centered) model can actually make sense in certain practical applications.

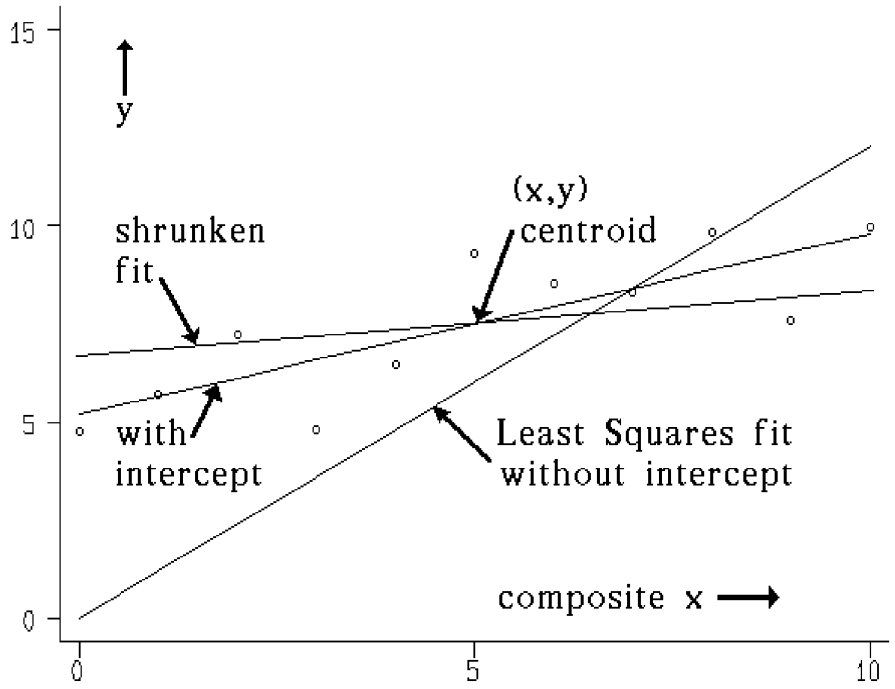
The only information not being used in the centered version of a model with no intercept is that expressed by the restriction $\overline{y} = \overline{x}^T\beta$; in models with an intercept, the difference $\overline{y} - \overline{x}^T\beta$ is simply the "implicit" intercept at $x = \vec{0}$, as described in Section §3.5 above. In fact, it is tempting to proceed as if $\overline{y} = \overline{x}^T\beta$ is merely a restriction on the length of the regression coefficient vector, β, that doesn't need to be addressed until we have reached the final, visual re-regression (VRR) phase of our analyses.

On the other hand, Figure 3.6 below illustrates a case where the information we ignored from the single degree-of-freedom for "no intercept" turns out to be rather traumatic once we have reached the final, VRR phase of our analysis. Here, both the shrinkage regression fit and the least-squares fit for a model with an intercept term are represented by a pair of lines which pass through $\overline{y}$ at $\overline{x}^T\beta^\bigstar$, where $\beta^\bigstar$ is the unit vector parallel to our favorite shrinkage regression estimator of β coefficients for the non-constant regressors. Unfortunately, neither of these two lines comes anywhere near to $y = 0$ at $x^T\beta^\bigstar = 0$. And yet only lines that actually do pass through $y = 0$ at $x^T\beta^\bigstar = 0$ are even candidates now that we have reached the final VRR phase of our analysis for a no-intercept model. Note that it would be truly unreasonable to simply translate (by moving parallel to itself) either one of these two fitted lines (down) so that it would pass through the (0, 0) origin; the fitted line would then again totally "miss" the data! My personal reaction in any sort of situation like that of Figure 3.6 would be that the available data cast very serious doubt on the cogency of a no-intercept model. After all, the least-squares line through (0, 0) would then yield mostly positive residuals to the left of the composite regressor mean value, $\overline{x}^T\beta^\bigstar$, and mostly negative residuals to the right of this same point.

On the other hand, the situation depicted in Figure 3.6 is somewhat pathological and, in many cases, the no-intercept least-squares fit will be a viable reference line for consideration during VRR. The equation for this least-squares fit is derived as follows: Let the $N \times 1$ vector of un-centered response values be $y^* = y + \overline{y}^* 1$, where y is centered; let the $N \times P$ matrix of un-centered, non-constant regressor coordinates be $X^* = X + 1 \cdot \overline{x}^{*T}$, where X is centered; and consider only rescaled estimates of the original shrinkage $b^\bigstar$ of the form $\widehat{\beta} = f \cdot \beta^\bigstar$. The sum-of-squares to be minimized is then $(y^* - X^*\widehat{\beta})^T(y^* - X^*\widehat{\beta})$, and the best rescaling is

$$f \;=\; \frac{y^{*T}X^{*}\beta^{\bigstar}}{\beta^{\bigstar T}X^{*T}X^{*}\beta^{\bigstar}} \;=\; \frac{(y^{T}X+N\cdot\bar{y}^{*}\cdot\bar{x}^{*T})\beta^{\bigstar}}{\beta^{\bigstar T}(X^{T}X+N\cdot\bar{x}^{*}\bar{x}^{*T})\beta^{\bigstar}}\;. \qquad \{\,3.12\,\}$$

Figure 3.6 Least Squares and Shrunken Fits For a Model with No Intercept



3.7 Shrinkage Residual Analyses

The residual vector, $r^{\bigstar}$, corresponding to the generalized shrinkage estimator, $b^{\bigstar}$, of equation { 3.1 } is

$$\begin{aligned} r^{\bigstar} \;&=\; y \,-\, X\,b^{\bigstar}, \\ &=\; (\,I \,-\, H\,\Delta\,H^{T}\,)\,y\,, \\ &=\; (\,I \,-\, 1\,1^{T}/N \,-\, H\,\Delta\,H^{T}\,)\,y\,, \end{aligned} \qquad \{\,3.13\,\}$$

where that last expression applies even when the response vector, y , hasn't been centered (so that $1^{T}\,y = 0$.)

Warning about possible confusions in notation: The $R \times 1$ vector of principal correlations of equation { 2.16 } is denoted by the symbol r , i.e. with no superscript. And, when individual elements of a correlation vector are referenced, they will have two subscripts; either r_{yj} for principal correlations or r_{yx_j} for marginal correlations, respectively. In contrast, the symbol $r^{\bigstar}$ (with a star superscript) represents an $N \times 1$ vector of fitted shrinkage-regression residuals. Note also that elements of the $r^{\bigstar}$ residual vector would be written with only one subscript ($r_i^{\bigstar}$ for the i-th observation.)

The shrinkage-regression residual vector can have a non-zero expected value even when $E(y \mid X) = X\beta$ of equation { 2.3 } is a correct expectation model because shrinkage estimators are usually biased estimators:

$$E(r^{\bigstar} \mid X) = (I - H \Delta H) X \beta = H (I - \Delta) \Lambda^{1/2} \gamma . \qquad \{ 3.14 \}$$

One sufficient condition for $E(r^{\bigstar} \mid X) = 0$ when the model is correct is that $\Delta = I$; after all, absolutely no shrinkage results in this extreme case (where $b^{\bigstar} = b^{o}$ and $r^{\bigstar} = r^{o}$ of ordinary least squares.)

The conditional variance-covariance matrix of the shrinkage-regression residual vector, given the observed regressor coordinates and under the assumption that $V(y \mid X) = \sigma^2 \cdot (I - 1\,1^T / N)$ of equation { 2.4 } is a correct dispersion model, is of the general form:

$$\begin{aligned} V(r^{\bigstar} \mid X) &= \sigma^2 \cdot (I - 1\,1^T / N - H \Delta H^T)^2 , \\ &= \sigma^2 \cdot (I - 1\,1^T / N - H H^T + H (I - \Delta)^2 H^T) . \\ &= V(r^{o} \mid X) + \sigma^2 \cdot H (I - \Delta)^2 H^T . \end{aligned} \qquad \{ 3.15 \}$$

3.7.1 Leverage Modifications Resulting from Shrinkage.

With x_i^T denoting the i-th row of the given regressor-coordinate X matrix, the shrinkage-regression prediction of the expected response at x_i^T would be

$$\text{estimated } E(y_i \mid x_i) = x_i^T b^{\bigstar} = h_i^T \Delta r \cdot \sqrt{y^T y} , \qquad \{ 3.16 \}$$

where h_i^T is the i-th row of the standard principal coordinates matrix, H of { 2.8 }, and r is the vector of principal correlations of equation { 2.16 }. The estimated variance of this prediction is

$$\text{estimated } V(y_i \mid x_i) = x_i^T V(b^{\bigstar} \mid X) x_i = \sigma^2 \cdot h_i^T \Delta^2 h_i , \qquad \{ 3.17 \}$$

where σ^2 is estimated by the least-squares residual-mean-square, $s^2 = RMS^{o}$.

Again, we define the leverage of the i-th observation on the regression, as in equation { 2.50 }, to be

$$\Lambda_i = \frac{\text{Predictive Variance}}{\text{Residual Variance}} = \frac{h_i^T \Delta^2 h_i}{[\,(N-1)/N - h_i^T h_i + h_i^T (I-\Delta)^2 h_i\,]} . \qquad \{ 3.18 \}$$

It is clear from equation { 3.18 } that shrinkage can only reduce the leverage of every regressor combination! After all, the numerator predictive variance is maximized and the denominator residual variance is minimized at the ordinary least squares solution; $\max \Lambda_i = h_i^T h_i / [\,(N-1)/N - h_i^T h_i\,]$ is achieved at $\Delta = I$.

When $h_i^T h_i$ is relatively large for the i-th observation, this means that x_i^T (the i-th row of X) is rather large as measured in the metric of $(X^T X)^{-1}$ and, thus, that the i-th regressor combination is relatively remote from the centroid of (possibly highly ill-conditioned) regressor coordinates. This remoteness gets translated by the least-squares fitting algorithm into a corresponding small residual variance, implying that the least-squares fit will be pulled relatively close to the corresponding observed response value, y_i.

When the shrinkage pattern is not uniform, some elements of Δ will be decreasing much more rapidly than others. For example, if $h_i^T = (\, h_{i1}, h_{i2}, \ldots h_{iR}\,)$ has some relatively large trailing coordinates [i.e. $|\, h_{iR}\,|, |\, h_{i(R-1)}\,|, \ldots$ are large relative to $|\, h_{i1}\,|, |\, h_{i2}\,|, \ldots$] and the shrinkage pattern utilizes declining deltas [$\delta_1 > \delta_2 > \ldots > \delta_R$], then the leverage of that regressor combination will decrease very quickly with shrinkage. After all, such an $h_i^T = (\, h_{i1}, h_{i2}, \ldots h_{iR}\,)$ gets most of its leverage from the minor-principal-axis dimensions that shrinkage is systematically de-emphasizing.

3.7.2 Standardized/Studentized Shrinkage Residuals.

The i-th residual is standardized by dividing it by the "usual" estimate of its standard deviation, the square root of the i-th diagonal element of { 3.15 } with σ^2 estimated by s^2 of { 2.35 } :

$$r_i^{\bigstar s} = \frac{r_i^{\bigstar}}{s \cdot \sqrt{(N-1)/N - h_i^T (\,2\Delta - \Delta^2\,) h_i}} . \qquad \{ 3.19 \}$$

These standardized residuals do not follow Student's-t distribution under normal theory because $r_i^{\bigstar}$ and s are not statistically independent. As in equation { 2.46 } , the estimator of σ^2 that is independent of $r_i^{\bigstar}$ is $s^2_{(-i)}$ of

$$(\, N - P - 2\,) \cdot s^2_{(-i)} = (\, N - P - 1\,)\, s^2 - (\, r_i^o\,)^2 / [(N-1)/N - h_i^T h_i\,] . \qquad \{ 3.20 \}$$

The i-th shrinkage residual is thus studentized as follows:

$$t_i^{\bigstar} = \frac{r_i^{\bigstar}}{s_{(-i)} \cdot \sqrt{(N-1)/N - h_i^T (2\Delta - \Delta^2) h_i}}, \qquad \{ 3.21 \}$$

$$= r_i^{\bigstar s} \cdot \sqrt{\frac{N-P-2}{[N-P-1-(r_i^{\bigstar s})^2]}}$$

The results derived here in Section §3.5 on analysis of shrinkage residuals can be summarized as follows:

> Shrinkage reduces the overall leverage of every regressor combination. But, when shrinkage is not uniform, the leverage of some observations may be reduced much more quickly than others. While shrinkage increases the average size of fitted residuals, some residuals may actually become smaller. In other words, shrinkage can change not only the relative magnitudes of fitted residuals but also of their standardized and studentized versions.

References for Chapter Three

Casella, G. (1980). "Minimax ridge regression estimation." **Annals of Statistics** 8, 1036-1056.

Casella, G. (1985). "Condition numbers and minimax ridge-regression estimators." **Journal American Statistical Association** 80, 753-758.

Crone, L. (1972). "The singular value decomposition of matrices and cheap numerical filtering of systems of equations." **Journal Franklin Institute** 294, 133-136.

Draper, N. R. and Smith, H. (1981). **Applied Regression Analysis**, Second Edition. New York: John Wiley.

Draper, N. R. and Van Nostrand, R. C. (1979). "Ridge regression and James-Stein estimation: review and comments." **Technometrics**, 21, 451-466.

Hoerl, A. E. (1962). "Application of ridge analysis to regression problems." **Chemical Engineering Progress** 58, 54-59.

Hoerl, A. E. and Kennard, R. W. (1970a). "Ridge regression: biased estimation for nonorthogonal problems." **Technometrics** 12, 55-67.

Hoerl, A. E. and Kennard, R. W. (1970b). "Ridge regression: applications to nonorthogonal problems." **Technometrics** 12, 69-82.

Hoerl, A. E. and Kennard, R. W. (1975). "A note on a power generalization of ridge regression." **Technometrics**, 17, 269.

Goldstein, M. and Smith, A. F. M. (1974). "Ridge-type estimators for regression analysis." **Journal Royal Statistical Society**, B, 36, 284-291.

Johnson, N. L. and Kotz, S. (1970). **Distributions in Statistics: Continuous Univariate Distributions-1**. (Chapter 17, Gamma Distribution, including "Chi Square.") New York, John Wiley.

Krishnamurthi, L. and Rangaswamy, A. (1987). "The equity estimator for marketing research." **Marketing Science** 6, 336-357.

Krishnamurthi, L. and Rangaswamy, A. (1989). "Response function estimation using the equity estimator." Warton Working Paper 89-030R, University of Pennsylvania.

Lowerre, J. M. (1974). "On the mean square error of parameter estimates for some biased estimators." **Technometrics**, 16, 461-464.

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

Massy, W. F. (1965). "Principal components regression in exploratory statistical research." **Journal American Statistical Association** 60, 234-256.

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

Obenchain, R. L. (1975b). "Ridge analysis following a preliminary test of the shrunken hypothesis." **Technometrics**, 17, 431-441. (Discussion: McDonald, G. C., 443-445.)

Obenchain, R. L. (1978). "Good and optimal ridge estimators." **Annals of Statistics**, 6, 1111-1121.

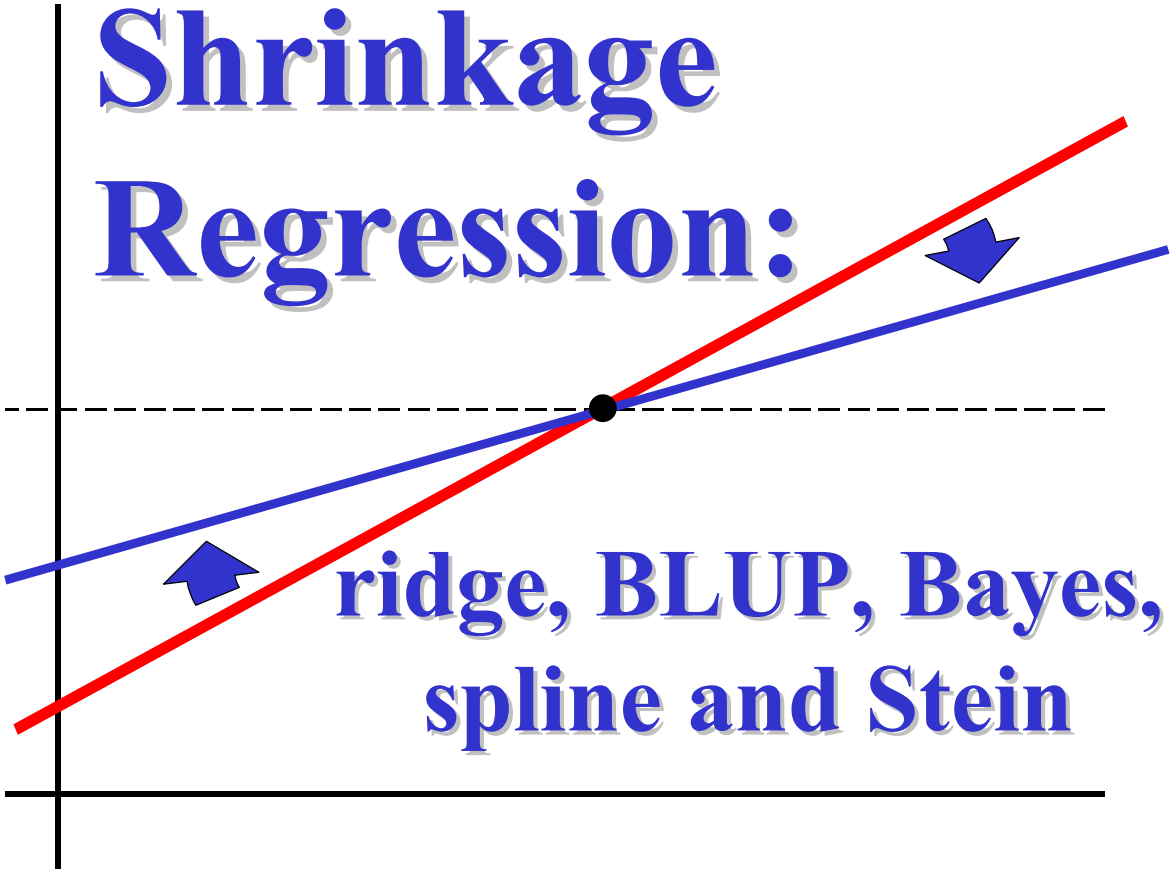
Piegorsch, W. W. and Casella, G. (1989). "The early use of matrix diagonal increments in statistical problems." **Siam Review** 31, 428-434.

Rao, C. R. (1973). **Linear Statistical Inference and its Applications,** 2nd edition. New York: John Wiley & Sons.

Strawderman, W. E. (1978). "Minimax adaptive generalized ridge regression estimators." **Journal American Statistical Association** 73, 623-627.

Theil, H. (1963). "On the use of incomplete prior information in regression analysis." **Journal American Statistical Association** 58, 401-414.

Thisted, R. (1976). "Ridge regression, minimax estimation, and empirical bayes methods." **Technical Report No. 28, Division of Biostatistics**, Stanford University.



Chapter 04: The Risk of Shrinkage

Bob Obenchain, Ph.D.
softR$_X$ freeware
13212 Griffin Run
Carmel, Indiana 46033-8835

Chapter 4: THE RISK OF SHRINKAGE

What criterion does one use to determine an ideal amount of shrinkage in a particular situation?

In attempting to address this question, we consider two "standard" setups for general linear models. In sections §4.1 to §4.3, the symbols β and σ^2 will denote the unknown, true values for the classical (fixed effect) regression coefficient vector and the residual variance, respectively. On the other hand, when mixed (fixed and random coefficient) models are considered in section §4.4, the expected value of β will be denoted by β_o while its variance will be denoted by Σ_β .

The common thread that binds the arguments presented in this chapter is that desirable forms/extents of shrinkage are characterized as using variance-bias tradeoffs to reduce measures of the overall mean-squared-error risk in estimating β. The major advantage of adopting the β, σ^2, β_o , Σ_β notation described above is that risk formulas can utilize these unknown quantities essentially as if they had known values.

On the other hand, the truly pivotal, simplifying assumption that we make here in all sections of Chapter 4, except section §4.3, is that the regression shrinkage factors (the multiplicative δ_i terms) are known constants. This allows us to view shrinkage estimators as linear estimators and, thus, to develop simple, closed-form expressions for statistical characteristics (bias, variance, mean-squared-error risk, etc.) of shrinkage estimators. This "non-stochastic" shrinkage formulation is of questionable cogency in addressing the "full range" of practical questions that can arise in actual applications of shrinkage regression to ill-conditioned data. But it does provide us with a good starting-off point. In fact, we will see that the problem of selecting a/the "best" type of shrinkage is really a rather difficult question to address even in our possibly oversimplified, non-stochastic shrinkage formulation.

The main distinction between the "optimal" shrinkage formulations of section §4.1 and the "good" shrinkage formulations of section §4.2 is that the corresponding measures of overall mean-squared-error risk are scalar-valued and matrix-valued, respectively. Scalar-valued risk measures can usually be unambiguously minimized, so the corresponding shrinkage estimators are called "optimal" in section §4.1. Matrix-valued measures of risk are realistically "multivariate," and "orderings" of risk matrices can be ambiguous. Thus the arguments of section §4.2 simply compare shrinkage estimators to the least-squares estimator, labeling a shrinkage estimator "good" when it dominates least-squares in <u>every</u> mean-squared-error sense for a specified β, σ^2 pairing.

Features that the optimal and good estimators described in sections §4.1 and §4.2 share include the following two fundamental disclaimers:

(i) Whether or not a given set of shrinkage factors, Δ , yield an optimal or good $b^{\bigstar}$ depends upon β and σ^2. No fixed, given Δ can yield an optimal or good $b^{\bigstar}$ for every β and σ^2.

(ii) One never knows in practical applications which $b^{\bigstar}$'s actually are optimal or good. Again, these properties depend upon the unknowns, β and σ^2, that are to be estimated from the data at hand.

Because the formulas that define optimal and good estimators fail to yield "operational" versions of those estimators, it is probably best to think of these concepts as simply defining target values for appropriate forms and extents of shrinkage. Identification of shrinkage estimators "most likely" to be optimal or good in a practical application is the primary topic of our next chapter, Chapter 5. And Chapters 6 through 9 outline a spectrum of alternative methods for identifying desirable patterns of shrinkage.

4.1 Classical "Optimal" Shrinkage

The classical mean-squared-error risk matrix of $b^{\bigstar} = G \Delta c$ as an estimator of the unknown, fixed β vector, where Δ is a given diagonal matrix of non-stochastic generalized shrinkage factors, is:

$$\text{MSE}(b^{\bigstar}) = \text{E}[(b^{\bigstar} - \beta)(b^{\bigstar} - \beta)^{\text{T}}] = G \, \text{MSE}(\Delta c) \, G^{\text{T}} , \qquad \{ 4.1 \}$$

where G is the $P \times R$ semi-orthogonal matrix of principal axis direction cosines for the centered regressors matrix, X , of equation { 2.8 } and

$$\text{MSE}(\Delta c) = \sigma^2 \, \Delta^2 \, \Lambda^{-1} + (I - \Delta) \, \gamma \, \gamma^{\text{T}} \, (I - \Delta) \qquad \{ 4.2 \}$$

is the mean-squared-error matrix of Δc as an estimator of the true vector of uncorrelated components, γ of equation { 2.19 }. Note that MSE (Δc) is the sum of two matrices, namely:

(i) the diagonal variance matrix, $\sigma^2 \, \Delta^2 \, \Lambda^{-1}$, which is void of covariance terms because the components of Δc are uncorrelated when Δ is non-stochastic,

plus

(ii) the rank-one matrix, $(I - \Delta) \, \gamma \, \gamma^{\text{T}} \, (I - \Delta)$, with squared bias terms along its diagonal and bias cross-product terms off that diagonal.

4.1.1 Diagonal Elements of Mean Squared Error Matrices

For our first specific example of a scalar-valued measure of risk, let us focus upon any single diagonal element, say the i-th, of the mean-squared-error matrix of $\Delta\, c$:

$$\mathrm{MSE}(\delta_i c_i) \;=\; \sigma^2 \delta_i^2/\lambda_i \;+\; (1-\delta_i)^2 \gamma_i^2 \;. \qquad \{\, 4.3 \,\}$$

Now $\mathrm{MSE}(\delta_i c_i)$ of { 4.3 } will clearly change as the i-th δ-factor changes. In fact, the partial derivative of $\mathrm{MSE}(\delta_i c_i)$ with respect to δ_i is

$$\partial \mathrm{MSE}(\delta_i c_i)/\partial \delta_i \;=\; 2\sigma^2 \delta_i/\lambda_i \;-\; 2(1-\delta_i)\gamma_i^2 \;, \qquad \{\, 4.4 \,\}$$

while the second partial derivative is a non-negative constant...

$$\partial^2 \mathrm{MSE}(\delta_i c_i)/\partial \delta_i^2 \;=\; 2\sigma^2/\lambda_i \;+\; 2\gamma_i^2 \;. \qquad \{\, 4.5 \,\}$$

It follows from { 4.5 } that equating $\partial \mathrm{MSE}(\delta_i c_i)/\partial \delta_i$ of { 4.4 } to zero will yield a MINIMUM value for $\mathrm{MSE}(\delta_i c_i)$ as long as either $\sigma^2 > 0$ or $\gamma_i^2 > 0$. This optimal amount of shrinkage for the i-th uncorrelated component, c_i, is

$$\begin{aligned} \delta_i^{\mathrm{MSE}} \;&=\; \gamma_i^2 \,/\, [\, \gamma_i^2 + (\, \sigma^2 \,/\, \lambda_i \,) \,] \;=\; \lambda_i \,/\, [\, \lambda_i + (\, \sigma^2 \,/\, \gamma_i^2 \,) \,] \,, \qquad \{\, 4.6 \,\} \\ &=\; \phi_i^2 \,/\, (\, \phi_j^2 \;+\; 1 \,) \;=\; (\, 1 + \phi_j^{-2} \,)^{-1}, \end{aligned}$$

where $\phi_j^2 = \gamma_i^2 \lambda_i/\sigma^2$ of { 2.24 } is the noncentrality parameter of the F-ratio, $F_i = c_i^2 \lambda_i/s^2$ of { 2.22 }, for testing the hypothesis that the i-th true uncorrelated component, γ_i, of β is zero. It follows that

$$\mathrm{MSE}(\delta_i c_i) \;\geq\; \sigma^2 \cdot \lambda_i^{-1} \cdot \delta_i^{\mathrm{MSE}}, \qquad \{\, 4.7 \,\}$$

with equality only at $\delta_i \;=\; \delta_i^{\mathrm{MSE}}$. Note, in particular, that the lower bound on $\mathrm{MSE}(\delta_i c_i)$ of { 4.7 } involves the first power of δ_i^{MSE}...not its square.

See Figures 4.2, 4.3, and 4.4 in section §4.2 of this chapter for graphs that show how $\mathrm{MSE}(\delta_i c_i)$ changes as δ_i decreases from 1 to 0 in the cases where $\phi_i^2>1$, $\phi_i^2=1$, or $\phi_i^2<1$, respectively. Technically, these figures actually display "relative" mean-squared-errors, defined as $\mathrm{MSE}(\delta_i c_i)/\mathrm{MSE}(c_i) = \lambda_i \cdot \mathrm{MSE}(\delta_i c_i) \,/\, \sigma^2$.

Next, note that δ_i^{MSE} of { 4.6 } can never be negative nor larger than one,

$$0 \;\leq\; \delta_i^{\mathrm{MSE}} \;\leq\; 1 \,. \qquad \{\, 4.8 \,\}$$

On the other hand, δ_i^{MSE} will be zero only when γ_i is zero and/or σ^2 is plus infinity. Similarly, δ_i^{MSE} will be one only when γ_i^2 is plus infinity and/or σ^2 is zero. But cases where either γ_i^2 or σ^2 is infinite or else σ^2 is zero are of little or no practical interest in applications. Therefore, the optimal shrinkage range of PRACTICAL INTEREST is

$$0 \leq \delta_i^{MSE} < 1 ,$$

with equality only when γ_i is zero.

4.1.2 MSE Measures Depending Only Upon Diagonal Elements

It so happens that several well known, scalar-valued measures of the "overall" mse risk in $b^{\bigstar} = G \Delta c$ depend only upon the diagonal elements of the MSE (Δc) matrix. And all such measures are obviously minimized when $\delta_i = \delta_i^{MSE}$ of { 4.6 } for i = 1, ..., R [where R = rank(X).] Two such examples of overall mse risk that depend only upon the diagonal elements of MSE(Δc) are...

(i) Summed Mean Squared Error [Hoerl and Kennard(1970a)]:

$$\begin{aligned} SMSE(b^{\bigstar}) &= \text{trace}\{ E [(b^{\bigstar} - \beta)(b^{\bigstar} - \beta)^T] \} , \\ &= \text{trace}\{ G\, MSE(\Delta c)\, G^T \} , \\ &= \text{trace}\{ MSE(\Delta c)\, G^T G \} = \text{trace}\{ MSE(\Delta c) \}, \end{aligned} \qquad \{ 4.9 \}$$

and

(ii) Summed, Scaled Predictive Mean Squared Error [Mallows(1973)]:

$$\begin{aligned} PMSE(b^{\bigstar}) &= 1 + (E \| Xb^{\bigstar} - X\beta \|^2) / \sigma^2 , \\ &= 1 + [\sum_{i=1}^{R} \lambda_i\, MSE(\delta_i c_i)] / \sigma^2 . \end{aligned} \qquad \{ 4.10 \}$$

4.1.3 Weighted Mean Squared Error Measures

So far, we have only considered scalar-valued risk-of-shrinkage criteria that ignore the off-diagonal bias cross-product terms in the MSE (Δc) matrix. Scalar-valued risk measures that may (or may not) depend upon off-diagonal bias terms are forms of "weighted" mean-squared-error:

$$\begin{aligned} wmse(b^{\bigstar}, W) &= E [(b^{\bigstar} - \beta)^T W (b^{\bigstar} - \beta)] , \\ &= \sigma^2 \cdot \text{trace}(M \Delta^2 \Lambda^{-1}) + \gamma^T (I - \Delta) M (I - \Delta) \gamma , \end{aligned} \qquad \{ 4.11 \}$$

where W and M are matrices of weights. W is a $P \times P$ non-stochastic weight matrix that is always taken to be symmetric and either non-negative definite (i.e. $\alpha^T W \alpha \geq 0$ for every α) or positive definite (i.e. $\alpha^T W \alpha > 0$ for every $\alpha \neq 0$.) Similarly, $M = G^T W G$ is the corresponding weight matrix for the uncorrelated components of $b^\star$.

Note that { 4.9 } and { 4.10 } are both special cases of weighted mse:

$$\text{wmse}(b^\star, I) = E [(b^\star - \beta)^T (b^\star - \beta)] = \text{SMSE}(b^\star) ,$$

for the positive definite choice $W = I$, and

$$\text{wmse}(b^\star, X^T X) = \sigma^2 [\text{PMSE}(b^\star) - 1] ,$$

for the rank R choice $W = X^T X$.

Consider the following result, from Obenchain(1978), that applies whenever the weight matrix, W, is positive definite...

> WEIGHTED MEAN-SQUARED-ERROR OPTIMALITY: If σ^2 is strictly positive, β is finite, and W is positive definite, then wmse($b^\star$, W) of { 4.11 } is minimized by choice of non-stochastic shrinkage factors Δ at:
>
> $$\delta_i = \gamma_i \cdot \lambda_i \cdot \eta_i / (\sigma^2 \cdot m_{ii}) , \qquad \{ 4.12 \}$$
>
> where η_i is the i-th element of $\eta = (D + M^{-1})^{-1} \gamma$, $M = ((m_{ij})) = G^T W G$, and D is the diagonal matrix with i-th diagonal element ϕ_i^2 / m_{ii}. Equivalently, $\Delta \gamma = D \eta = D (D + M^{-1})^{-1} \gamma$.

Note from the second expression for wmse($b^\star$, W) in { 4.11 } that

$$\partial \text{wmse} / \partial \delta_i = 2 (\sigma^2 m_{ii} / \lambda_i) \cdot \delta_i - 2 \sum_{j=1}^{R} m_{ji} \gamma_i \gamma_j \cdot (1 - \delta_j) , \qquad \{ 4.13 \}$$

$$\partial^2 \text{wmse} / \partial \delta_i \partial \delta_j = 2 \, m_{ji} \gamma_i \gamma_j \quad \text{for } j \neq i,$$

and

$$\partial^2 \text{wmse} / \partial \delta_i^2 = 2 \, m_{ii} (\sigma^2 / \lambda_i + \gamma_i^2) .$$

The conditions that $\sigma^2 > 0$ and that W be positive definite (so that $m_{ii} > 0$) assure that the second derivatives matrix $((\partial^2 \text{wmse} / \partial \delta_i \partial \delta_j))$ will be positive definite and, therefore, that the MINIMUM wmse will occur at $\partial \text{wmse} / \partial \delta_i = 0$ for i = 1,...,R. If we define $\eta \equiv M (I - \Delta) \gamma$, then $\partial \text{wmse} / \partial \Delta = 0$ yields the “fixed point" version of equations { 4.12 }; in other words, the optimal Δ is defined in terms of an η vector that is, in turn, a function of this Δ .

When $\gamma_i = 0$, the corresponding optimal δ_i is clearly also zero, even from the fixed-point form of { 4.12 }. Any nonzero component must be finite because β is finite, so we can then

multiply { 4.12 } by γ_i , yielding $\delta_i\gamma_i = D_{ii}\eta_i$ where $D_{ii} = \phi_i^2/m_{ii}$ is the i-th diagonal element of the diagonal D matrix introduced above. As a result $\eta \equiv M (I - \Delta) \gamma$ can be rewritten as $\eta = M \gamma - M D \eta$ or $(I + M D) \eta = M \gamma$. Of course, $(I + M D)^{-1} = (D + M^{-1})^{-1} M^{-1}$ whenever M is positive definite, so we arrive at the closed-form solution $\eta = (D + M^{-1})^{-1}\gamma$ that we sought.

OBSERVATIONS ON OPTIMALLY WEIGHTED MEAN-SQUARED-ERROR:

Equation { 4.12 } shows that the optimal amount of wmse shrinkage will always be $\delta_i = \delta_i^{MSE}$ of { 4.6 } whenever the weight matrix, W, is such that $M = G^T W G$ is diagonal. After all, the terms in the summation of { 4.13 } with $j \neq i$ vanish when M is diagonal, and the two remaining terms both contain a m_{ii} factor that then cancels out!

Another special case where the optimal amount of wmse shrinkage will always be $\delta_i = \delta_i^{MSE}$ of { 4.6 } occurs when only one component, say γ_k , of γ is nonzero. And this statement holds for every positive definite choice of weight matrix, W . All terms in the summation of { 4.13 } with either $i \neq k$ or $j \neq k$ again vanish, leaving only the $-(m_{kk}\gamma_k^2) \cdot 2(1 - \delta_k)$ term in the k-th equation. Of course, $\delta_i^{MSE} = 0$ for $i \neq k$ in this case !

Once we have established some interesting results about mean-squared-error in specific directions of space in the next subsection, §4.1.4, we will show how weighted mse using the one-parameter family of weight matrices $W = I + (\zeta\text{-}1)\beta\beta^T$ provides some profound new insights about definitions of mean-squared-error optimal shrinkage.

4.1.4 The Mean Squared Error in Specific Directions

When α is a fixed vector of unit length (i.e. $\alpha^T\alpha = 1$), the rank-one weight matrix $W = \alpha\alpha^T$ yields...

$$\text{wmse}(b^\star, \alpha\alpha^T) = \alpha^T \text{MSE}(b^\star) \alpha = \text{MSE}(\pm \alpha^T b^\star) , \qquad \{ 4.14 \}$$

$$= \sigma^2 \xi^T\Delta^2\Lambda^{-1}\xi + [\xi^T(I - \Delta)\gamma]^2 .$$

MSE($\alpha^T b^\star$) measures the size of the component in the mean-squared-error of $b^\star$ in the direction parallel to $\pm\alpha$ in P-dimensional Euclidean space or, equivalently, the mean-squared-error of the linear combination $\alpha^T b^\star$ as an estimator of $\alpha^T \beta$. The last expression in { 4.14 } follows by writing $\xi^T = \alpha^T G$ to denote the unit vector that expresses the orientation of α relative to the principal axes of the centered regressors matrix, X .

Some additional results from Obenchain(1978) apply here...

> DIRECTIONAL MEAN-SQUARED-ERROR OPTIMALITY: If σ^2 is strictly positive, β is finite, and α is a fixed vector of unit length, then:

(i) MSE($\alpha^T b^\star$) does not depend upon δ_i whenever $\xi_i \equiv \alpha^T g_i = 0$, where g_i is the direction-cosine vector of the i-th principal axis of the centered regressors matrix.

(ii) MSE($\alpha^T b^\star$) depends upon δ_i whenever $\xi_i \equiv \alpha^T g_i \neq 0$ and, in fact, is minimized by choice of non-stochastic shrinkage factor, δ_i , at:

$$\delta_i = \delta_i(\alpha) = \alpha^T \beta \gamma_i \lambda_i / [\xi_i (\sigma^2 + \sum\nolimits^* \gamma_j^2 \lambda_j)] , \qquad \{ 4.15 \}$$

where $\sum^*$ denotes summation only over subscripts j such that $\xi_j \neq 0$.

The above results are demonstrated, first, by noting that $\xi^T = \alpha^T G$ implies that the i-th element of ξ will be $\xi_i \equiv \alpha^T g_i$. Thus the component of a generalized shrinkage estimator, $b^\star$, in the $+\alpha$ direction is $\alpha^T b^\star = \xi^T \Delta c = \sum \xi_j \delta_j c_j$. Thus, when an element of the ξ vector is null, there can be NO dependency of the component of $b^\star$ in the $\pm \alpha$ direction upon that δ_i factor. As a result, the corresponding "directional" mean-squared-error will also NOT depend in any way upon that δ_i factor.

It follows from { 4.14 } that

$$\partial MSE/\partial \delta_i = 2 (\sigma^2 \xi_i^2/\lambda_i) \cdot \delta_i - 2 [\xi^T(I - \Delta)\gamma] \cdot \xi_i \gamma_i , \qquad \{ 4.16 \}$$

$$\partial^2 MSE/\partial \delta_i \partial \delta_j = 2 \xi_i \gamma_i \xi_j \gamma_j \quad \text{for } j \neq i,$$

and

$$\partial^2 MSE/\partial \delta_i^2 = 2 \xi_i^2 (\sigma^2/\lambda_i + \gamma_i^2) .$$

The conditions that $\sigma^2 > 0$ and $\xi_i \neq 0$ assure that the relevant $((\partial^2 MSE/\partial \delta_i \partial \delta_j))$ sub-matrix will be positive definite and, therefore, that the MINIMUM directional MSE will occur at $\partial MSE/\partial \delta_i = 0$ for i = 1,...,R. As anticipated above, $\partial MSE/\partial \delta_i = 0$ reduces simply to $0 = 0$ and, thus, provides NO INFORMATION whenever $\xi_i = 0$. Thus the $\partial MSE/\partial \delta_i = 0$ equations for 1 = 1,...,R can be summarized as requiring that:

each optimal δ_i be proportional to $\gamma_i \lambda_i / \xi_i$ whenever its $\xi_i \neq 0$.

Furthermore, the common constant-of-proportionality is $k = [\xi^T(I - \Delta)\gamma]/\sigma^2$, which is an expression that depends upon these optimal δ_i shrinkage factors. Note, however, that we can rewrite k as: $k\sigma^2 = \xi^T\gamma - \sum \delta_j \xi_j \gamma_j = \xi^T\gamma - k\sum^* \gamma_j^2 \lambda_j$. Since $\alpha^T\beta = \xi^T\gamma$, the optimal k is thus $k(\alpha) = \alpha^T\beta / (\sigma^2 + \sum^* \gamma_i^2 \lambda_i)$ as in { 4.15 }.

A shrinkage formula equivalent to { 4.15 } requires, for i=1, ..., R, that

$$\delta_i \xi_i \gamma_i = \phi_i^2 \frac{\sum \xi_j \gamma_j}{(1 + \sum^* \phi_j^2)} \quad \text{whenever } \xi_i \neq 0.$$

OBSERVATIONS ABOUT OPTIMAL DIRECTIONAL MEAN-SQUARED-ERROR:

It is immediately clear from the form of equation { 4.15 } that the optimal shrinkage factors for the $-\alpha$ direction are identical to those for the $+\alpha$ direction.
Equation { 4.15 } also shows that the optimal amount of MSE shrinkage for direction $\alpha = g_i$ will always be: $\delta_i = \delta_i^{MSE}$ of { 4.6 } along the i-th axis. And, furthermore, δ_j will be completely undetermined for all axes $j \neq i$ when $\alpha = g_i$ because MSE($\delta_i c_i$) does not depend in any way upon shrinkage factors along orthogonal directions.

When $\beta = 0$, every α is orthogonal to β in the sense that $\alpha^T\beta = 0$, and { 4.15 } shows that drastic shrinkage ($\Delta = 0$) is optimal for all choices of α in this case. Even when $\beta \neq 0$, there still is a (R-1) dimensional space of α's that are orthogonal to β , and drastic shrinkage ($\Delta = 0$) is again optimal by { 4.15 } for all of these directions, α , such that $\alpha^T\beta = 0$. In other words,...

> Shrinking the least squares solution all of the way to ZERO by taking $\Delta = 0$, assures that NO ERRORS WHATSOEVER will be made in any direction orthogonal to the unknown, true β .

While it might be reassuring to have this knowledge of the desirability of drastic shrinkage along directions orthogonal to β, we don't want to make egregious errors in that one (unknown) direction that happens to be parallel to β.

When $\beta \neq 0$, taking α parallel to β yields $\xi_i = \gamma_i/\sqrt{\gamma^T\gamma}$. Note that no optimal shrinkage factor along the i-th principal axis, δ_i , is defined for minimizing MSE parallel to β whenever $\gamma_i = 0$. But, when $\gamma_i \neq 0$, optimal shrinkage for minimizing MSE parallel to β is necessarily of the form

$$\delta_i = k^{(=)}\lambda_i \quad \text{for} \quad k^{(=)} = (\gamma^T\gamma) / (\sigma^2 + \gamma^T\Lambda\gamma) . \qquad \{ 4.17 \}$$

In fact, it follows that MSE[$\beta^T b^\bigstar / \sqrt{\beta^T\beta}$] $\geq \sigma^2 k^{(=)}$ with equality only at $\Delta = k^{(=)} \Lambda$ of { 4.17 }.

Equation { 4.17 } provides the following, almost astounding insight ! The generalized shrinkage estimator with $\Delta = k^{(=)}\Lambda$ is:

$$b^\bigstar = k^{(=)} G \Lambda c = k^{(=)} X^T y ,$$

where $X^T y$ is the $(P \times 1)$ vector of inner products between the centered regressor matrix and the centered response vector. In other words,...

> Although the true β is unknown, we know that the generalized shrinkage estimator achieving minimal mean-squared-error parallel to β has the SAME RELATIVE MAGNITUDES as does the vector of MARGINAL INNER PRODUCTS of the regressors with the response.

Furthermore, these marginal relative magnitudes are generally different from those of the least-squares coefficients ...unless regressors are uncorrelated as in { 2.7 }.

One curious property of the optimal factors $\Delta = k^{(=)}\Lambda$ of { 4.17 } is that some of them may exceed 1. As we shall see in Chapter 6, the normal-theory maximum likelihood estimator of $k^{(=)}$ is $\sum r_{yi}^2 \lambda_i^{-1}/[R^2+(1-R^2)/n]$, where the r_{yi} are the principal correlations of { 2.23 }. Thus the true values of the optimal $\delta_i = k^{(=)}\lambda_i$ and also their normal-theory estimators are both potentially quite sensitive to the eigenvalue spectrum, Λ, of regressors; the largest and smallest λ's cause two different sorts of sensitivity to ill-conditioned regressors!

In subsection §4.1.5 we will seek a balance between the conflicting objectives of minimizing mean squared error parallel to and orthogonal to the unknown, true β. But let us first comment on some of the LESS intuitively satisfying implications of equation { 4.15 }. Specifically, for directions, α, that are "oblique" not only to the principal axes of the given, centered regressors matrix, X, but also "oblique" to the true coefficient vector, β, it turns out that the optimal shrinkage factors of { 4.15 } may NOT fall within the range $0 \leq \delta_i(\alpha) \leq 1$. By "oblique" here I simply mean that angles between directions are neither zero nor an exact multiple of 90°. Anyway, we are NOT talking about "trivial" violations of the $0 \leq \delta_i(\alpha) \leq 1$ range restriction that arise simply because principal axis i is strictly orthogonal to the chosen α; shrinkage factor δ_i remains undetermined by { 4.15 } in these cases, and δ_i could take on any numerical value without making any real difference in MSE($\alpha^T b^\star$).

To be specific, consider the special case of MSE(β_1+β_2) where $\beta = \gamma$ (i.e. G = I), $\alpha^T = (+1, +1, 0^T)/\sqrt{2}$, $\lambda_1 = \lambda_2$, and the first two components of β are of the form $\beta_1 \equiv f \cdot \beta_2$ ...where $\beta_2 \neq 0$ and the constant factor, f, is not 0, +1, or -1. In this case, relationship { 4.15 } determines only the first two shrinkage factors, δ_1 and δ_2. And the values for these factors that minimize MSE($b_1^\star$+$b_2^\star$) are...

$$\delta_1 = \frac{(f+1)\cdot f}{(a^2+f^2+1)} \text{ and } \delta_2 = \frac{(f+1)}{(a^2+f^2+1)} ,$$

where $a^2 = \sigma^2/\beta_2^2$. Note that the cases we have explicitly excluded give "reasonable" answers, namely...

f= -1 implies that α is orthogonal to β, and the optimal $\delta_1 = \delta_2$=0,
f=0 implies $\beta_1 = 0$, $\delta_1 = \delta_1^{MSE} = 0$, and δ_2=δ_2^{MSE},
and f=+1 implies the nonzero components of α lie parallel to their β components and $\delta_1 = \delta_2$=2/(a^2+2).

But the "bad news" is that...

f more negative than -1 gives a negative optimal value for δ_2,
f within (-1, 0) gives a negative optimal value for δ_1,
f within $0.5 \pm \sqrt{0.25 - a^2}$ yields an optimal δ_2 larger than +1 when a^2<0.25,

and f larger than $1+a^2$ gives an optimal δ_1 larger than +1.

Results of the above sort are, intuitively speaking, really not very pleasing or insightful. Apparently, equation { 4.15 } can exploit highly specialized [and potentially "weird"] forms of shrinkage patterns (like the $\delta_1 = f \cdot \delta_2$ pattern in the above example were $\beta_1 \equiv f \cdot \beta_2$) to gain reductions in mean-squared-error. Unfortunately, the kind of "information" exploited in these special cases is unrealistic in the sense that it is either "not available" or is potentially "incorrect/misleading" in most actual applications of shrinkage regression to real data.

It was probably quite "intrepid" of us to even consider the problem of optimizing a scalar valued function, $MSE(\alpha^T b^\bigstar)$, by choice of a relatively "large" number of shrinkage parameters, δ_1, δ_2, ..., δ_R, when the fundamental parameters involved (β and σ^2) are actually unknowns. So let us be content here with the relatively simple and intuitive results we obtained for the special cases where α corresponds to a principal regressor axis or is either strictly orthogonal to or strictly parallel to β.

4.1.5 Balancing Components of MSE Risk Parallel To and Orthogonal To the Unknown True Coefficient Vector

Any generalized shrinkage estimator can be written as

$$b^\bigstar = b^{(=)} \cdot \beta + b^{(\perp)}, \qquad \{ 4.18 \}$$

where the scalar $b^{(=)}$ determines the size of the component of $b^\bigstar$ parallel to β and $b^{(\perp)}$ is the component of $b^\bigstar$ orthogonal to β. To display explicit formulas, let β^+ denote the row-vector defining the Moore-Penrose inverse of the column-vector of regression coefficients, β. Thus $\beta^+ = 0$ when $\beta = 0$, and otherwise $\beta^+ = \beta^T / \beta^T\beta$.

Orthogonal projection in Euclidean space is accomplished by a linear operator that can be represented as a symmetric, idempotent, and uniquely determined matrix, Rao(1973) pp 46-47. Thus...

$\beta\beta^+$ is the projection for the space, of dimension 1 or 0, parallel to β,

and

$I - \beta\beta^+$ is the projection for the space, of dimension P-1 or P, orthogonal to β.

Thus equation { 4.18 } implies that

$$b^{(=)} = \beta^+ b^\bigstar = \gamma^+ \Delta c, \qquad \{ 4.19 \}$$

and

$$b^{(\perp)} = (I - \beta\beta^+) b^\bigstar = G (I - \gamma\gamma^+) \Delta c . \qquad \{ 4.20 \}$$

It would clearly be desirable for generalized shrinkage estimators based upon non-stochastic Δ factors to have the features that $b^{(=)}$ tends to be close to 1, at least when $\beta \neq 0$, and that $b^{(\perp)}$ tends to be small.

What non-stochastic choice of Δ in { 4.19 } makes $b^{(=)}$ a minimum mean-squared-error estimator of ONE ? Clearly, no choice of Δ can be of any real help if $\beta = \gamma = 0$; $b^{(=)} \equiv 0$ in this case. So suppose that $\gamma \neq 0$, and note that we then have

$$MSE(\gamma^{+}\Delta\, c) = E[(\gamma^{+}\Delta\, c - 1)^2] \; = \{\sigma^2\gamma^T\Delta^2\Lambda^{-1}\gamma + [\gamma^T(I-\Delta)\gamma]^2\} / (\gamma^T\gamma)^2 \,, \qquad \{ 4.21 \}$$

where we used the fact that $1 = \gamma^{+}\gamma \; = \gamma^T\gamma/\gamma^T\gamma$. Now { 4.21 } is identical to { 4.14 } except, of course, that γ may not be of unit length like ξ . However, the mse optimal shrinkage factors are again $\Delta = k^{(=)}\Lambda$ for $k^{(=)} = (\gamma^T\gamma) / (\sigma^2 + \gamma^T\Lambda\gamma)$ as in { 4.17 }, where the minimum value of { 4.21 } thereby attained is $\sigma^2/(\sigma^2 + \gamma^T\Lambda\gamma)$. Therefore, our only new insight so far is that $k^{(=)}\gamma^{+}\Lambda\, c$ is, when viewed as a known linear function of the uncorrelated components, c, a minimum mean-squared-estimator of $\gamma^{+}\gamma$, which is either 1 or 0. [It's perhaps unfortunate we restricted attention to estimators that have to depend upon the sample estimate, c , of γ ; if we hadn't, we might have found even "better" estimates of 0 and 1 ..., namely, 0 and 1 !]

What non-stochastic choice of Δ in { 4.20 } makes $b^{(\perp)}$ as small as possible ? Well $\Delta = 0$ clearly provides the global minimum! Unfortunately, this choice makes $MSE(b^{(=)}) = 1$ when $\beta \neq 0$, which can be considerably larger than the minimum, $\sigma^2/(\sigma^2 + \gamma^T\Lambda\gamma)$, attained at $\Delta = k^{(=)}\Lambda$, or the value $\sigma^2\gamma^T\Lambda^{-1}\gamma/(\gamma^T\gamma)^2$, attained at the unbiased, least-squares solution, $\Delta = I$.

Therefore let us now consider the problem of minimizing the weighted mean-squared-error orthogonal to β, $wmse(b^{\bigstar}, I - \beta\beta^{+})$, under a restriction on the amount of mean-squared-error allowed parallel to β, $wmse(b^{\bigstar}, \beta\beta^{+})$. The optimal solution will, again, obviously be $\Delta = 0$ when $\beta = 0$. So suppose that $\beta \neq 0$, and note that we can then write the Lagrange multiplier equation for the constrained optimization problem as...

$$\Psi(\Delta) \; = wmse(b^{\bigstar}, I - \beta\beta^{+}) \; + \; \zeta \cdot [\, wmse(b^{\bigstar}, \beta\beta^{+}) + \eta^2 \; - \; U \,] \,, \qquad \{ 4.22 \}$$

where ζ is the Lagrange multiplier, η^2 is a slack variable, and U is the desired upper limit on $wmse(b^{\bigstar}, \beta\beta^{+})$. The range of interest for U will be $\sigma^2(\gamma^T\gamma) / (\sigma^2 + \gamma^T\Lambda\gamma) \leq U \leq \sigma^2(\gamma^T\gamma)$, where the lower limit is achieved at $\Delta = k^{(=)}\Lambda$ and the upper limit is achieved at $\Delta = 0$. As usual, $\partial\Psi/\partial\eta = 0$ gives us the familiar condition that the optimum must occur where either the slack is zero [$\eta = 0$] or the multiplier is zero [$\zeta = 0$]; and $\partial\Psi/\partial\zeta = 0$ gives us the familiar condition that the constraint must be satisfied [$wmse(b^{\bigstar}, \beta\beta^{+}) \leq U$] .

Let us denote the minimal value of $wmse(b^{\bigstar}, I - \beta\beta^{+})$ achievable in { 4.22 } by choice of Δ subject to the restriction $wmse(b^{\bigstar}, \beta\beta^{+}) \leq U$ by the function h(U). Then $\partial\Psi/\partial U = 0$ implies

$$\partial h(U)/\partial U \; = \; -\zeta \;. \qquad \{ 4.23 \}$$

The larger is U, the more Δ can deviate from $k^{(=)}\Lambda$ and effect a reduction in h(U), yet still satisfy wmse($b^{\bigstar}$, $\beta\beta^{+}$) $\leq$ U .

Consider, now, the special case where $\zeta = 1$ in { 4.23 }. Slack must clearly be zero for any optimum that occurs here because $\zeta \neq 0$. Thus the primary message of { 4.23 } is that an EXACT BALANCE is struck at $\zeta = 1$ between the rate of change (decrease or increase) in weighted mean-squared-error orthogonal to β , h(U), and the rate of change (increase or decrease) in mean-squared-error parallel to β , U .

Now that we have gained the fundamental insights provided by equations { 4.18 } through { 4.23 }, we can skip over a great deal of other technical details [such as establishing the convexity of the minimization problem in { 4.21 }] by noting that our equation { 4.12 } actually provides an almost complete solution to the optimal tradeoffs problem. After all, the combined (orthogonal plus parallel) weight matrix in the Lagrange equation, namely...

$$W = I + (\zeta - 1) \cdot \beta\beta^{+} ,$$

will be positive definite as long as $\zeta > 0$. And we also already know that $\Delta = 0$ is optimal for $\zeta = 0$. Therefore, we can summarize all of our findings as follows:

Choice of the ζ hyper-parameter is critical in establishing tradeoffs between the mean-squared-error component parallel to β and those orthogonal to β . Small numerical values of ζ emphasize reductions in mean-squared-error orthogonal to β ; large numerical values of ζ emphasize reduction in mean-squared-error parallel to β .
In the limit as ζ approaches zero, there is effectively no constraint on the amount of mean-squared-error allowed parallel to β, and the optimal SHRINKAGE TARGET is declared to be $\Delta = 0$. This strategy would be REALLY EASY to implement in actual practice, but it is clearly also a rather EXTREME strategy.

In the limit as ζ approaches plus infinity ($+\infty$), there is effectively no constraint on the amount of mean-squared-error allowed orthogonal to β, and the optimal SHRINKAGE TARGET is declared to be the $\Delta = k^{(=)}\Lambda$ of { 4.17 }. The resulting shrinkage $b^{\bigstar}$ will be parallel to the marginal inner products vector, $X^{T}y$. This strategy would require an estimate of $k^{(=)}$ to be derived from the data. (Alternatively, one might simply adjust the length of this $b^{\bigstar}$ to maximize its likelihood of being β , i.e. minimize the resulting residual sum-of-squares.) But this, too, would be a rather EXTREME strategy.

The choice $\zeta = 1$ establishes EQUILIBRIUM in the tradeoff between mean-squared-error components orthogonal to β and the mean-squared-error parallel to β . And ζ=1 corresponds to the weight matrix $W = I$ in { 4.23 }. As a result, the optimal SHRINKAGE TARGET values are the $\delta_{i}^{MSE} = \phi_{i}^{2} / (\phi_{j}^{2} + 1)$ of { 4.6 }. This strategy is relatively difficult to implement in actual practice because it could require estimates for all R of the canonical signal-to-noise ratios, i.e. the $\phi_{j}^{2} = \gamma_{i}^{2}\lambda_{i}/\sigma^{2}$ noncentrality parameters of { 2.24 }. But, of the three strategies discussed here, this also appears to be the MOST REASONABLE overall strategy.

4.1.6 Canonical Form for Optimal Shrinkage of a Single Fixed-Effect Estimator

Because the different components of a vector of regression coefficients can be of different numerical sizes and can have different variances, consider the possibility of rescaling each individual component [using the appropriate unknown, multiplicative constant] so that the variance of its least-squares estimate will equal one. Once placed in this canonical form, we will see that the rescaled size of each fixed-effect component plays a pivotal role.

The j-th uncorrelated component, γ_j , of β is placed in its canonical form by dividing it by its standard deviation, $\sigma \cdot \lambda_j^{-1/2}$. An additive – error model for this rescaled component is thus of the form:

$$\begin{matrix}\text{FIXED-EFFECT}\\ \text{ESTIMATE}\end{matrix} = \begin{matrix}\text{FIXED-EFFECT}\\ \text{SIGNAL}\end{matrix} + \begin{matrix}\text{STANDARDIZED}\\ \text{NOISE}\end{matrix} ,$$

where the standardized noise has mean zero and variance one. Note that $\phi_j = \gamma_j \cdot \lambda_j^{1/2} / \sigma$ is then both the size of the fixed-effect signal and the expected value of the fixed-effect estimate. In other words, the rescaled component has variance one and expected value $\phi_j = \gamma_j \cdot \lambda_j^{1/2} / \sigma$. Note also that the optimal extent of shrinkage for this canonical fixed-effect component is $\delta_i^{MSE} = \phi_i^2 / (\phi_j^2 + 1)$, as in { 4.6 }.

Canonical forms will be used at the end of Section §4.4 to display an exact analogy between the fixed-effect and random-effect formulations for optimal shrinkage. These canonical forms will also be used extensively in the risk simulations of Chapter §6.

4.2 Classical "Good" Shrinkage

Matrix-valued mean-squared-error risk criteria are much more in keeping with the primary theme of our book than are scalar-valued criteria. After all, we always prefer to stress our theme: Simultaneous estimation of 2 or more regression coefficients is nothing less than a full-blown problem in multivariate analysis.

Swindel and Chapman(1973) originally defined “good" ridge estimators only within the one-parameter ridge family of Hoerl and Kennard(1970a). But we will apply their definition to all generalized shrinkage estimators as follows:

> **GOOD SHRINKAGE ESTIMATORS:** A generalized shrinkage estimator, $b^{\bigstar} = G \Delta c$ of { 3.1 } with non-stochastic shrinkage factors Δ , will be said to be GOOD for a specified β , σ^2 pairing if and only if it dominates the least-squares estimator $b^o \equiv X^+ y = G c$ in EVERY mean-squared-error sense.

To explore mathematically equivalent formulations, consider the following...

The EXCESS Mean-Squared-Error of b^o relative to $b^\bigstar$ is defined to be simply the corresponding algebraic difference in $P \times P$ Mean-Squared-Error matrices (the least-squares variance matrix minus a shrinkage estimator MSE matrix.) This difference can be thought of as depending only upon the form and extent of shrinkage applied to $b^\bigstar$; in fact, we will commonly think of this difference as being primarily a function of the diagonal matrix of shrinkage factors, Δ . In any case, we will denote this difference matrix by

$$\text{EMSE}(b^\bigstar) \;=\; \text{MSE}(b^o) \;-\; \text{MSE}(b^\bigstar) \;=\; G\, \text{EMSE}(\Delta c)\, G^T , \qquad \{ 4.24 \}$$

where G is the $P \times R$ semi-orthogonal matrix of principal axis direction cosines for the centered regressors matrix, X , of equation { 2.8 } and

$$\text{EMSE}(\Delta c) \;=\; \sigma^2 (I \;-\; \Delta^2) \Lambda^{-1} \;-\; (I - \Delta) \gamma \gamma^T (I - \Delta) . \qquad \{ 4.25 \}$$

Three equivalent conditions that assure that the generalized shrinkage estimator $b^\bigstar = G \Delta c$ with non-stochastic shrinkage factors Δ is GOOD for a specified β , σ^2 pairing can now be stated as follows:

(i) $\text{EMSE}(b^\bigstar)$ is a positive definite matrix, or { 4.26 }

(ii) $\text{MSE}(\alpha^T b^o) > \text{MSE}(\alpha^T b^\bigstar)$ for every unit vector α , or { 4.27 }

(iii) $\text{wmse}(b^o, W) > \text{wmse}(b^\bigstar, W)$ for every { 4.28 }

nonzero, non-negative definite weight matrix, W. The equivalence of { 4.26 } and { 4.27 } follows immediately from the very definition of a positive-definite matrix. Theobald(1974), Theorem 1, claimed it was sufficient to consider only positive definite matrices, W, in { 4.28 }. But, in a 1979 letter to me, Masashi Okamoto showed that the stronger condition that all nonzero, non-negative definite weight matrices be considered is necessary to imply equivalence with the other two definitions.

The highly specialized form of the EMSE matrix of { 4.25 } [namely a diagonal matrix minus a symmetric, rank-one matrix], turns out to imply some key results about GOOD shrinkage estimators. We will need the following lemma from Obenchain(1978)...

> **OBENCHAIN'S LEMMA:** If $D = \text{Diag}(d_1, d_2, ..., d_p)$ is a $p \times p$ positive-definite diagonal matrix and $z^T = (z_1, z_2, ..., z_p)$ is a row vector with $p > 2$ elements, then the length of the z-vector is critical in determining whether or not matrices of the general form $A \;=\; D (I \;-\; z z^T) D$ are positive definite. Specifically,
>
> (i) A will be positive definite iff $z^T z < 1$,

(ii) A will be non-negative definite of rank $(p-1)$ iff $z^T z = 1$, and

(iii) A will have $(p-1)$ positive eigenvalues and 1 negative eigenvalue iff $z^T z > 1$.

Furthermore, the eigenvector, τ, of A corresponding to a eigenvalue, λ, is of the general form:

(a) $\tau =$ (i-th column of the identity matrix) and $\lambda = d_i^2$ if $z_i = 0$ for some i, and

(b) $\tau \propto (D^2 - \lambda \cdot I)^{-1} D z$ if $\lambda \neq d_i^2$ for any $i = 1, 2, ..., p$.

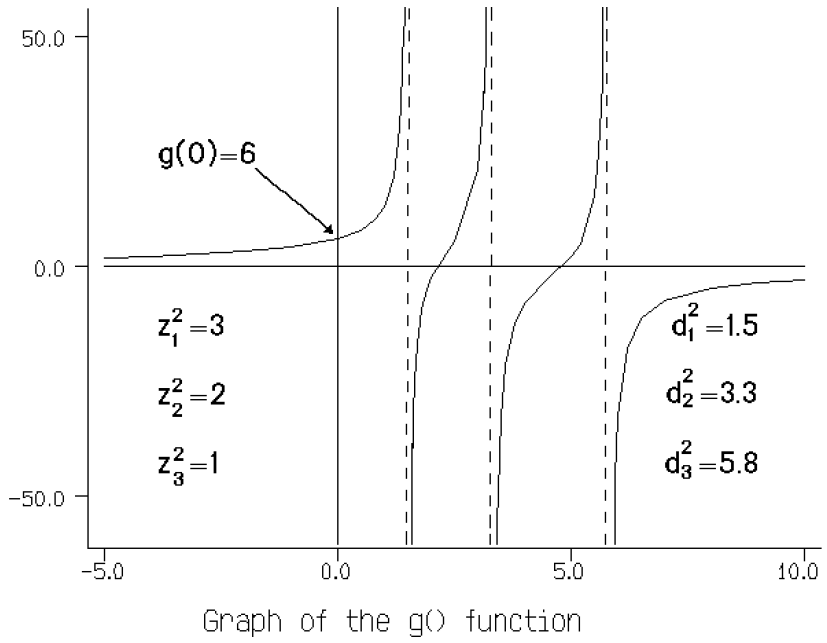
Since $\tau \neq 0$ is to be an eigenvector of A with eigenvalue λ, we know that $A\tau = D^2\tau - D z$ $z^T D \tau$ must be of the special form $\lambda \cdot \tau$. We can rewrite this condition as $(D^2 - \lambda \cdot I) \tau = k \cdot D z$ where k is the scalar $k = z^T D \tau$. The problem of finding a complete set of eigenvectors and eigenvalues for A becomes simple [as in case (a), above] when $z = 0$, so suppose that $z \neq 0$. Furthermore, if $\lambda \neq d_i^2 > 0$ for i=1, 2, ..., p, then $(D^2 - \lambda \cdot I)$ will be invertible, and τ will necessarily be of the special form $\tau = k \cdot (D^2 - \lambda \cdot I)^{-1} D z$ of case (b) above, at least when k is non-zero. But this means that $k = z^T D \tau = k \cdot z^T D (D^2 - \lambda \cdot I)^{-1} D z = k \cdot g(\lambda)$, where g denotes the scalar valued function

$$g(\lambda) = \sum_{i=1}^{p} z_i^2 d_i^2 (d_i^2 - \lambda)^{-1}. \qquad \{ 4.29 \}$$

It follows that each eigenvalue of A must either coincide with one of the positive d_i^2 values or else be a solution of $g(\lambda) = 1$, which excludes the possibility that $k = z^T D \tau = 0$ in case (b).

Letting d_{MIN}^2 denote the smallest numerical value of d_i^2 for which a corresponding $z_i^2 > 0$, it is clear that $g(\lambda) \geq 0$ and is strictly increasing on $-\infty < \lambda < d_{MIN}^2$ and, furthermore, that $g(\lambda) = 1$ has exactly one solution in this interval. After all, if this minimal solution to $g(\lambda) = 1$ is denoted by λ_{MIN} , then it cannot be a multiple eigenvalue of A because the eigenspace of $(D^2 - \lambda_{MIN} \cdot I)^{-1} D z$ is clearly of rank one. It follows that the numerical value of g() at $\lambda = 0$, namely $g(0) = z^T z$, is critical in determining whether the smallest eigenvalue of A is negative, zero, or positive. Specifically, $z^T z < 1$ implies that λ_{MIN} is strictly positive because $g(\lambda)$ has not yet reached the critical numerical value of 1 at $\lambda = 0$. Similarly, $z^T z = 1$ implies $\lambda_{MIN} = 0$, and $z^T z > 1$ implies $\lambda_{MIN} < 0$, as was to be shown.

Figure 4.1: g(λ) Function Numerical Example



In applying **OBENCHAIN'S LEMMA** to identify "good" shrinkage estimators, the following function will play the role of g(λ) of { 4.29 }.

RIDGE FUNCTION: The following scalar valued function of the generalized shrinkage factors, Δ, is called the Ridge Function:

$$\text{RF}(\Delta) = \sum_{j=1}^{R} \phi_j^2 \cdot (1 - \delta_j)/(1 + \delta_j), \qquad \{ 4.30 \}$$

$$= \sum_{j=1}^{R} [\delta_j^{MSE} \cdot (1 - \delta_j)] / [(1 - \delta_j^{MSE}) \cdot (1+\delta_j)],$$

where $\phi_j^2 = \gamma_j^2 \lambda_j / \sigma^2$ is, again, the unknown noncentrality of the F-ratio for the hypothesis that $\gamma_j = 0$ [i.e. the hypothesis that the j-th true component of β is zero.]

We can now state a theorem which is a mild generalization of the main result of Obenchain(1978)...

RIDGE FUNCTION THEOREM: If the parameters β and σ^2 of a classical, fixed effects linear model are such that $\beta^T\beta < \infty$ and $0 < \sigma^2 < \infty$, the given matrix of centered regressor coordinates X is of rank $R \geq 1$, and the generalized shrinkage factors Δ are non-stochastic on the range $0 \leq \delta_i < 1$ for i = 1, 2, ..., R, then

(i) the (R-1) largest eigenvalues of EMSE($b^{\bigstar}$) will always be positive,

(ii) the smallest eigenvalue of EMSE($b^{\bigstar}$) will also be positive iff $RF(\Delta) < 1$,

(iii) the smallest eigenvalue of EMSE($b^{\bigstar}$) will be zero iff $RF(\Delta) = 1$, and

(iv) the smallest eigenvalue of EMSE($b^{\bigstar}$) will be negative iff $RF(\Delta) > 1$. In this case, the eigenvector corresponding to the negative eigenvalue, ξ_R, has elements of the general form:

$$\alpha_i \propto \sum_{j=1}^{R} [g_{ij} \cdot (1 - \delta_j) \cdot \gamma_j] / [\, \sigma^2 \lambda_j (1 - \delta_j^2) + |\xi_R| \,] , \qquad \{ 4.31 \}$$

for i = 1, 2, ..., P, which defines the **INFERIOR DIRECTION** of P-dimensional space along which MSE($b^{\bigstar}$) exceeds MSE(b^{o}).

It will only be necessary to show that EMSE($b^{\bigstar}$) can be rewritten in a form to which OBENCHAIN'S LEMMA applies. But EMSE($b^{\bigstar}$) of { 4.24 } is $G\,A\,G^T$ where A of { 4.25 } is of the desired form with $D = (I - \Delta^2)^{1/2} \Lambda^{-1/2} \sigma$ and $z = \Lambda^{1/2} (I - \Delta^2)^{-1/2} (I - \Delta) \gamma / \sigma$. Note, in particular, that RF(Δ) of { 4.30 } is then $z^T z$, and that the α vector of { 4.31 } is necessarily of the form $G \tau$ for an eigenvector of A specified by case (b) of OBENCHAIN'S LEMMA because its eigenvalue is negative.

COMMENTS ON GOOD SHRINKAGE ESTIMATORS:

The numerical value attained by $RF(\Delta)$ is critical in determining the matrix mean-squared-error characteristics of generalized shrinkage estimators, $b^{\bigstar}$. Specifically, $RF(\Delta) < 1$ implies that, if the corresponding $b^{\bigstar}$ differs from b^{o}, this $b^{\bigstar}$ is “good" in the sense that it dominates b^{o} in ALL mean-squared-error senses.

On the other hand, if the ridge function EXCEEDS one, then there is at most one direction in P-dimensional space along which $b^{\bigstar}$ has larger mean squared error than does b^{o}.

In the one-parameter “ordinary ridge" family of Hoerl and Kennard(1970a), the shrinkage factors are restricted to be of the form $\delta_i = \lambda_i / (\lambda_i + k)$ for i = 1, 2, ..., p where k is a non-negative scalar. In this case, the ridge function is of the special form $RF(k) = \sum \phi_i^2 / (1 + 2\lambda_i k^{-1})$ and the main results of Swindel and Chapman(1973) follow as a special case of part (ii) of the ridge function theorem. Namely, every positive k value yields a good ordinary ridge estimator if $\sum \phi_i^2 < 1$; otherwise, the good range is $0 < k < 2 / |\eta_p|$ where η_p is the negative eigenvalue of $(X^T X)^{-1} - \beta\beta^T / \sigma^2$. As a result, the sufficient condition of Theobald(1974), Theorem 2, that $0 < k < 2\sigma^2 / \beta^T \beta$ tends to be much more stringent than is necessary.

THE (2/R)THS RULE-OF-THUMB: P, the number of (non-constant) predictor variables in our regression equation, is an upper bound for R = rank(X). Obenchain(1978) described a "(2/P)ths" guideline for GOOD shrinkage under the assumption that R = P. When R is less that P, my original quideline is more accurately described as a "(2/R)ths rule." Consider the problem of limiting shrinkage along each of the R principal regressor axes so that each of the R terms in { 4.30 } will not exceed 1/R. This is certainly one way of guaranteeing that the ridge function will not exceed one. This sufficient condition for "good"ness implies, for axis j, that

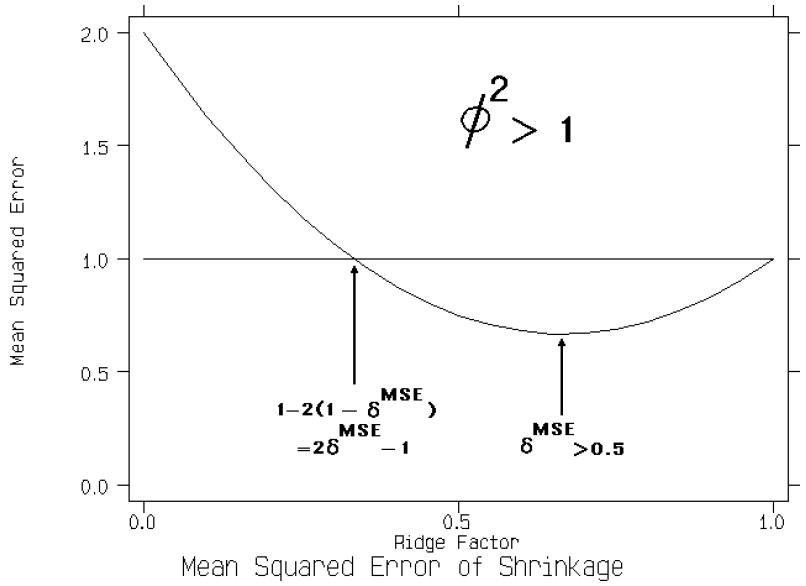
$$\delta_j^{MSE} \cdot (1 - \delta_j) \;\leq\; (1 - \delta_j^{MSE}) \cdot (1+\delta_j) / R$$

or, equivalently,

$$\delta_i \;\geq\; 1 \;-\; 2\,(1 - \delta_i^{MSE}) \cdot [\; 1 + (R - 1)\delta_i^{MSE} \;]^{-1}.$$

A set of sufficient conditions that are somewhat weaker, at least when R > 1, can thus be written as $\delta_i \geq 1 - 2\,(1 - \delta_i^{MSE}) / R$ for i=1,...,R, which means that the "good" shrinkage range will always be AT LEAST (2/R)ths of the "optimal" shrinkage range.

Figure 4.2: "Good" Range for Phi-Squared Greater Than 1.



When the rank of X is R=1, $\delta_1 c_1$ will be a "good" estimator of $\beta_1=\gamma_1$ even for shrinkage extents as much as TWICE the "optimal" extent; the good range will be $\delta_1^{MIN} \leq \delta_1 < 1$, where $\delta_1^{MIN} = \max(\, 0,\, 2 \cdot \delta_1^{MSE} - 1\,)$, as shown in Figures 4.2-4.4.

Figure 4.3: "Good" Range when Phi-Squared Equals 1.

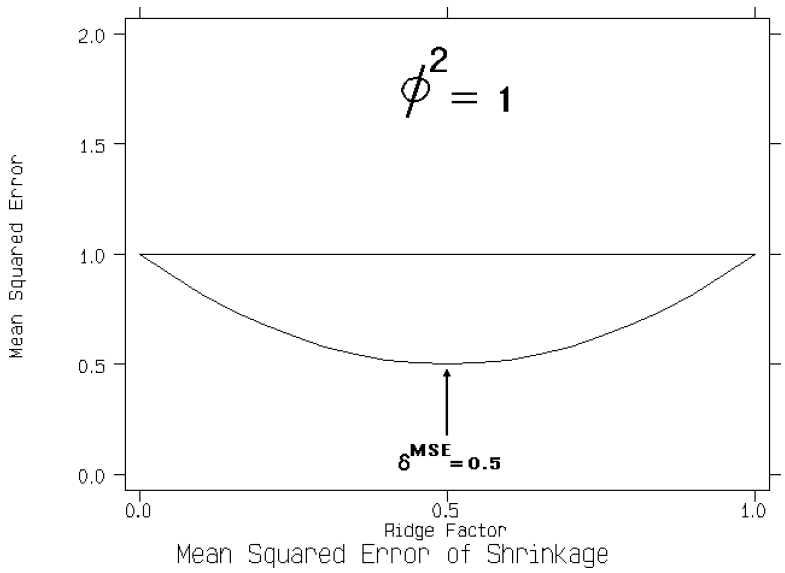
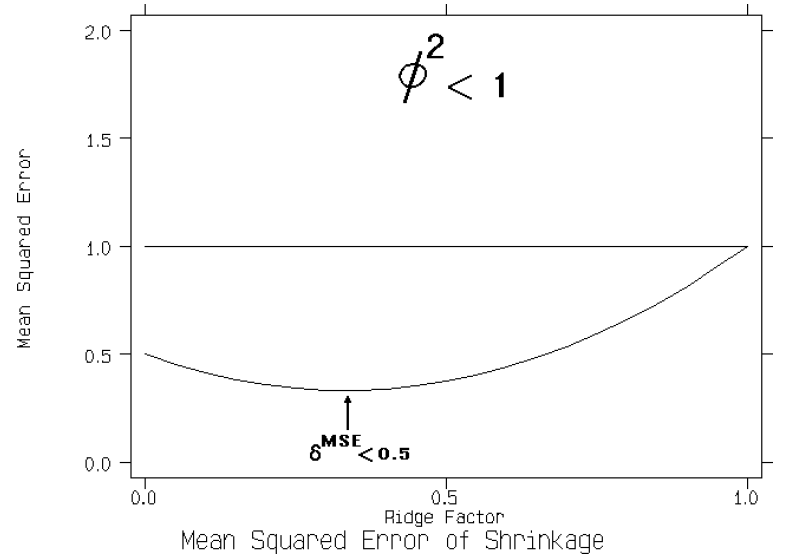


Figure 4.4: "Good" Range for Phi-Squared Less Than 1.



When the rank of X is two, RF(Δ^{MSE}) cannot exceed 1, and the good shrinkage range it thus at least $\delta_1^{MSE} \leq \delta_1 < 1$ and $\delta_2^{MSE} \leq \delta_2 < 1$.

When the rank of X exceeds two, RF(Δ^{MSE}) can exceed 1.

When the overall EXTENT of shrinkage is measured on the "multicollinearity allowance" scale, MCAL = $P - \delta_1 - \delta_2 - ... - \delta_R$, then the (2/R)ths Rule-of-Thumb for shrinkage can be quite simply stated as:

The "good" shrinkage range always includes AT LEAST (2/R)ths of the "optimal" range, namely...

$$0 \leq \text{MCAL} \leq 2 \cdot \text{MCAL}^{\text{MSE}} / R .$$

4.3 Classical "Ultimate" Shrinkage

Unlike the arguments we have explored so far here in the first two sections of Chapter 4, let us now specifically address the question: "What TARGET values should we use when shrinkage-factors are specifically recognized as being stochastic?"

After all, when we begin a new application of regression, we usually do not know in advance exactly how much of exactly which kind of shrinkage we might end up using. When our approach is "classical," the observed regressor values (the centered X coordinates) are assumed given, and we wish to make appropriate inferences about the conditional mean and variance of the distribution of responses, y , given those X realizations. When we "examine" the observed response values (using ridge trace displays and/or maximum likelihood calculations) to decide upon a form and extent for shrinkage, we end up using shrinkage factors that depend upon the observed y values. Therefore, the shrinkage factors we usually end up using are stochastic (given X .)

Our target values for stochastic shrinkage could still be the δ_i^{MSE} factors of { 4.6 }, of course, and I personally feel that the shrinkage target should be unchanged. However, let us first examine an alternative target.

What shrinkage factors would result from equating a generalized shrinkage estimator, $b^{\bigstar} = G \Delta c$, with the vector of true (fixed) coefficients, $\beta = G \gamma$? In other words, instead of merely reducing MSE risk (expected quadratic loss) of estimation, this choice would actually achieve zero loss! The equations that result state: $\Delta c = \gamma$. Therefore, whenever the i-th uncorrelated component of b^o is nonzero ($c_i \neq 0$), the corresponding ultimate choice for the i-th shrinkage factor would be:

$$\delta_i^{\text{ULT}} = \gamma_i / c_i . \qquad \{ 4.32 \}$$

Note that a strictly positive contribution to quadratic loss, $(\delta_i^{\text{ULT}} c_i - \gamma_i)^2$, can then occur only when an uncorrelated component of b^o happens to be zero, $c_i = 0$, while the corresponding true component is nonzero, $\gamma_i \neq 0$. Of course, the probability of observing an exactly null component, $c_i = 0$, is zero for any continuous (non-atomic) probability distribution of regression disturbance terms.

The formula for ultimate shrinkage, { 4.32 }, results in a stochastic choice for each ridge shrinkage factor. After all, the denominator c_i terms are assumed to be random given the observed regressor coordinates, X , so the δ_i^{ULT} factors are then also random.

The δ_i^{ULT} factors of { 4.32 } are clearly highly desirable target values. One might argue, perhaps, that the "natural" estimate of δ_i^{ULT} would be 0 when one believes that $\gamma_i = 0$ and 1 otherwise. On the other hand, an actual δ_i^{ULT} value could, of course, be negative! And no finite lower or upper bounds can be placed upon potential realizations of δ_i^{ULT} factors! In other words, the δ_i^{ULT} factors can represent sign-changes and/or expansions of the uncorrelated components of b^o rather than shrinkages. For example, it can be shown that the distribution of δ_i^{ULT} under normal-theory is bimodal with a less-likely, negative mode and a more-likely, positive mode at $2 / \sqrt{1 + (8/\phi_i^2)}$.

Vinod(1976) argued that the δ_i^{ULT} factors of equation { 4.32 } provide a "global minimum" in mean-squared-error of estimation. Actually, Vinod(1976) expressed results in terms of "additive eigenvalue inflation konstants," k_i . In this notation, shrinkage factors are of the form $\delta_i = \lambda_i / (\lambda_i + k_i)$, which is the generalization of equation { 3.6 } in which a potentially different amount is added to each eigenvalue of X^TX ; see Hoerl and Kennard(1970), sections §5 and §7. Using this notation, Vinod(1976) called

$$k_i^{MSE} = \sigma^2/\gamma_i^2 \qquad \{ 4.33 \}$$

"suboptimal" compared to

$$k_i^{ULT} = (c_i - \gamma_i) \cdot \lambda_i / \gamma_i \, . \qquad \{ 4.34 \}$$

In his response, Kennard(1976) said that { 4.34 } (or { 4.32 }) expresses a simple tautology: zero risk (or loss) can only be achieved when the γ_i components essentially have known values. In fact, Kennard points out that Vinod's "recommendation is just that of using the parameter value as its estimate."

In summary, equations { 4.32 } and { 4.34 } both essentially say... "Perform the exactly correct adjustment to c_i so that $\delta_i \cdot c_i$ will coincide exactly with γ_i ." Unfortunately, no advice on how to actually accomplish anything like this has been (or can be) given. After all, a stochastic target is (by its very definition) a constantly moving target!

> Actual reductions in RISK (over at least some parts of parameter space) can result from "aiming" at the fixed (but unknown) minimal MSE shrinkage target, { 4.6 } . Specific examples of this type are given in Chapter 6.
>
> No systematic reduction in LOSS has ever been demonstrated to be possible in any even remotely realistic situation.

4.4 Random Coefficient Shrinkage

While a general discussion of "mixed" linear models (models that contain both fixed and random regression coefficients) is best postponed until Chapter 7, we now discuss the generalization of some of the fixed-effect results of Section §4.1 to the case where β is random with expected value β_o and variance-covariance matrix Σ_β .

In the following discussion, it is quite important to remember that generalized shrinkage vector, $b^{\bigstar} = G\,\Delta\,c$, is still to be viewed as an estimator of the unknown, random β vector rather than as an estimator of the expectation vector, β_o. The uncorrelated components of the least squares estimator are still defined here to be $c = G^T b^o$, but the corresponding true components, γ, will now be random with expected value vector $E(\gamma) = \gamma_o \equiv G^T \beta_o$ and variance-covariance matrix $V(\gamma) = \Sigma_\gamma \equiv G^T \Sigma_\beta G$. The arguments leading to equation { 4.2 } of Section §4.1 still hold as long as the resulting mean-squared-error matrix is viewed as being conditional given a specific realization for γ :

$$MSE(\Delta c \mid \gamma) = \sigma^2 \Delta^2 \Lambda^{-1} + (I - \Delta)\,\gamma\,\gamma^T (I - \Delta). \qquad \{ 4.35 \}$$

The corresponding unconditional mean-square-error matrix is then of the general form

$$MSE(\Delta c) = \sigma^2 \Delta^2 \Lambda^{-1} + (I - \Delta)[\gamma_o \gamma_o^T + \Sigma_\gamma](I - \Delta), \qquad \{ 4.36 \}$$

in which the second term is no longer necessarily of rank one.

Two extreme, special cases of { 4.36 } will be of primary interest to us...

First of all, when $\Sigma_\beta = 0$ (so that $\beta \equiv \beta_o$ and $\gamma \equiv \gamma_o$) all random-effect risk measures revert to their somewhat more simple fixed-effect forms.

Secondly, the completely random coefficients case results when $\beta_o = 0$ and $\gamma_o = 0$.

The weighted mean-squared-error measure of equation { 4.11 } and the directional mean-squared-error measure of equation { 4.14 } take on the following forms when coefficients are random:

$$wmse(b^{\bigstar}, W) = \sigma^2 \cdot \text{trace}(M\Delta^2\Lambda^{-1}) + \gamma_o^T(I - \Delta)M(I - \Delta)\gamma_o + \text{trace}[\Sigma_\gamma(I - \Delta)M(I - \Delta)] \qquad \{ 4.37 \}$$

and

$$wmse(b^{\bigstar}, \alpha\alpha^T) = \sigma^2 \xi^T [\Delta^2\Lambda^{-1} + (I - \Delta)(\gamma_o \gamma_o^T + \Sigma_\gamma)(I - \Delta)]\xi, \qquad \{ 4.38 \}$$

where, again, $M = G^T W G$ and $\xi^T = \alpha^T G$. The corresponding generalizations of equations { 4.13 } and { 4.16 } for risk partial derivatives are straightforward, but closed form solutions like those of { 4.12 } and { 4.15 } for the shrinkage factors that minimize weighted or directional risks of random-coefficient shrinkage estimates are not obvious, except in certain special cases.

4.4.1 Shrinkage Risk of a Single Random Coefficient

Suppose that we start with an unbiased estimate, c , of an unknown, scalar-valued effect, γ, that has variance σ^2. In other words, given γ and σ^2, the conditional moments of c are E(c | γ, σ) = γ and V(c | γ, σ) = σ^2. [Note that, in the notation of section §4.1.1, the variance of the i-th uncorrelated component of the least-squares estimator would be written as σ^2/λ_i; here, that variance is simply being called σ^2.]

Next suppose that σ^2 has a fixed, unknown value while γ is random. Specifically, suppose that the expected value of γ is γ_o and its variance is σ_γ^2 :

$$E(\gamma) = \gamma_o \quad \text{and} \quad V(\gamma) = \sigma_\gamma^2 . \qquad \{ 4.39 \}$$

With δ denoting a known, non-stochastic shrinkage factor value, what are the mean-squared-error properties of $\delta \cdot c$ as an estimator of γ ? We again stress that we are viewing $\delta \cdot c$ as an estimator of γ itself, which is random when $\sigma_\gamma^2 > 0$, rather than as an estimator of γ_o, the fixed-effect, expected value of γ. The mean-squared-error of interest is thus

$$\begin{aligned} MSE(\delta \cdot c) &= E[(\delta \cdot c - \gamma)^2] \\ &= \delta^2 \cdot E(c^2) + E(\gamma^2) - 2 \cdot \delta \cdot E(c \cdot \gamma) \\ &= \delta^2 \cdot E(c^2) + (1 - 2 \cdot \delta) \cdot E(\gamma^2) \\ &= \delta^2 \cdot [\gamma_o^2 + \sigma_\gamma^2 + \sigma^2] + (1 - 2 \cdot \delta) \cdot [\gamma_o^2 + \sigma_\gamma^2] \\ &= \delta^2 \cdot \sigma^2 + (1 - \delta)^2 \cdot [\gamma_o^2 + \sigma_\gamma^2] . \qquad \{ 4.40 \} \end{aligned}$$

Now MSE($\delta \cdot c$) of { 4.40 } clearly changes as the δ-factor changes. In fact, the partial derivative of MSE($\delta \cdot c$) with respect to δ is

$$\partial MSE(\delta \cdot c) / \partial \delta = 2 \cdot \sigma^2 \cdot \delta - 2 \cdot (1 - \delta) \cdot [\gamma_o^2 + \sigma_\gamma^2] , \qquad \{ 4.41 \}$$

while the second partial derivative is a non-negative constant...

$$\partial^2 MSE(\delta \cdot c) / \partial \delta^2 = 2 \cdot [\sigma^2 + \gamma_o^2 + \sigma_\gamma^2] . \qquad \{ 4.42 \}$$

Equation { 4.42 } implies that equating $\partial MSE(\delta \cdot c) / \partial \delta_i$ of { 4.41 } to zero will yield a MINIMUM value for MSE($\delta \cdot c$) as long as $\sigma^2 > 0$ or $\gamma_o^2 > 0$ or $\sigma_\gamma^2 > 0$. This optimal amount of shrinkage is

$$\begin{aligned} \delta^{MSE} &= (\gamma_o^2 + \sigma_\gamma^2) / (\gamma_o^2 + \sigma_\gamma^2 + \sigma^2) , \qquad \{ 4.43 \} \\ &= \phi^2 / (\phi^2 + 1) = (1 + \phi^{-2})^{-1}, \end{aligned}$$

where $\phi^2 = (\gamma_o^2 + \sigma_\gamma^2) / \sigma^2$.

Again, the extreme cases of { 4.43 } are of special interest to us...

The fixed-effect results of equations { 4.3 } through { 4.6 } correspond to the special case of { 4.40 } through { 4.43 } where $\sigma_\gamma^2 = 0$ (so that $\gamma \equiv \gamma_o$.) Here $\phi^2 = \gamma_o^2 / \sigma^2$ is the unknown noncentrality parameter of the F-statistic for testing $\gamma_o = 0$ of { 2.22 } and { 2.23 }.

The completely random coefficient case results when $\gamma_o \equiv 0$. In this special case, $\phi^2 = \sigma_\gamma^2 / \sigma^2$ is an unknown, true ratio of variances, while an F-statistic is the corresponding ratio of sample variances.

Of course, we may also find ourselves in an "intermediate" situation where both $\sigma_\gamma^2 > 0$ and $\gamma_o \neq 0$. But the risk simulations of Chapter §5 will at least pin-down the extremes.

4.4.2 Canonical Form for Optimal Shrinkage of a Single, Completely-Random Effect

By dividing each component of a random coefficient vector by its noise standard deviation, we can place random-coefficient estimation problems in a canonical form analogous to that of Section §4.1.6 for fixed-effect models. The resulting "relative" standard deviation, $\phi = \sigma_\gamma / \sigma$ then plays a pivotal role.

An additive – error model for a rescaled random-effect estimate would be:

$$\begin{matrix}\text{RANDOM-EFFECT} \\ \text{ESTIMATE}\end{matrix} = \begin{matrix}\text{RANDOM-EFFECT} \\ \text{SIGNAL}\end{matrix} + \begin{matrix}\text{STANDARDIZED} \\ \text{NOISE}\end{matrix} ,$$

where the standardized noise has mean zero and variance one. Note that the random-effect signal has mean zero and variance ϕ^2 , while the random-effect estimate has mean zero and variance $\phi^2 + 1$. Note also that the optimal extent of shrinkage for this canonical random-effect would be $\delta_i^{MSE} = \phi^2 / (\phi^2 + 1)$, as in { 4.43 }.

4.5 Summary

In this chapter, we have used a wide variety of rather technical arguments to address an extremely important practical issue, that of selecting TARGET VALUES for shrinkage in regression models. The first-time reader may well ask "What do all of those theorems and special cases have to say about what to do in general, shrinkage regression practice?" Here are my personal opinions...

The δ^{MSE} generalized shrinkage factors, defined as in either { 4.6 } or { 4.43 }, seem to be the target values that make the most sense from the widest selection of alternative points-of-view. These factors establish optimal variance-bias tradeoffs that minimize a wide variety of univariate measures of mean-squared-error.

Once one adopts a truly multivariate (maxtix-valued risk) point-of-view, one still probably wishes to investigate shrinkage path shapes that lead generally "toward" (if not exactly "through") Δ^{MSE} on their way from $\Delta = I$ to $\Delta = 0$. However, a cautious practitioner might well wish to stop well short of Δ^{MSE} as his/her conservative choice for an extent of shrinkage. Objective shrinkage practitioners may ultimately find that the most important concepts introduced in this chapter are: (i) the "inferior direction" associated with excessive shrinkage, equation { 4.31 }, [as well as its associated excess-MSE eigenvalue spectrum] and (ii) the "(2/R)ths Rule-of-Thumb" given at the end of section §4.2.

In both our fixed-effect and random-coefficient formulations, shrinkage results from multiplying an unbiased estimator by a non stochastic factor, δ, on the range $0 \leq \delta \leq 1$. Bias is introduced when $\delta < 1$, but the corresponding variance is thereby reduced by a multiplicative factor of δ^2. A lower bond on the MSE risk (variance plus squared-bias) associated with this shrinkage results from multiplying the variance of the unbiased estimator by δ ; potential minimum risk decreases linearly with δ, { 4.7 }. On the other hand, this lower limit on risk is actually achieved only when applying the "right" extent of shrinkage, $\delta = \delta^{MSE}$ of { 4.6 } or { 4.43 }.

There is an exact analogy between the fixed-effect and completely-random-effect formulations for optimal shrinkage. The ϕ parameter takes the form of either a standardized fixed-effect when $\sigma_\gamma = 0$,

$$\phi = \gamma / \sigma = \text{(expected signal) / (standard deviation of additive noise)},$$

or a ratio of standard deviations when $E(\gamma) = 0$,

$$\phi = \sigma_\gamma / \sigma = \text{(standard deviation of signal) / (standard deviation of additive noise)}.$$

In fact, the optimal shrinkage target (in both extreme and all intermediate cases) is always of the general form $\delta^{MSE} = \phi^2 / (\phi^2 + 1)$ for $\phi^2 = (\gamma^2 + \sigma_\gamma^2) / \sigma^2$.

References for Chapter Four

Hoerl, A. E. and Kennard, R. W. (1970). "Ridge regression: biased estimation for non orthogonal problems." **Technometrics**, 12, 55-67.

Kennard, R. W. (1976). Letter to the Editor. **Technometrics**, 18, 504-505.

Obenchain, R. L. (1978). "Good and optimal ridge estimators." **Annals of Statistics** 6, 1111-1121.

Okamoto, M. (1979). Personal communication.

Rao, C. R. (1973). **Linear Statistical Inference and Its Applications, Second Edition**. New York: John Wiley and Sons.

Swindel, B. F. and Chapman, D. D. (1973). "Good ridge estimators." Abstracts Booklet, New York Joint Statistical Meetings, page 126.

Theobald, C. M. (1974). "Generalizations of mean square error applied to ridge regression." **Journal Royal Statistical Society B**, 36, 103-105.

Vinod, H. D. (1976). Letter to the Editor. **Technometrics**, 18, 504.

Further Reading for Chapter Four

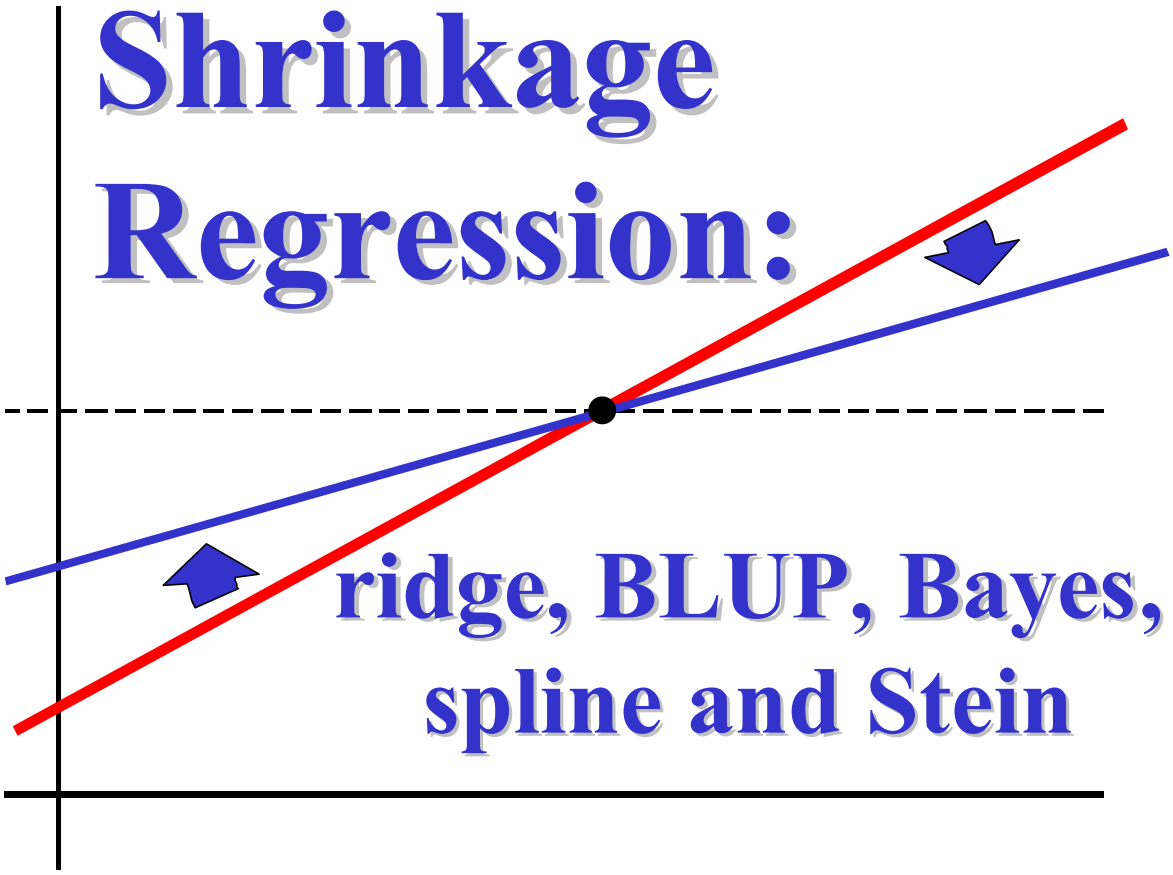
Farebrother, R. W. (1975). "The minimum mean square error linear estimator and ridge regression." **Technometrics**, 17, 127-128.

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

Farebrother, R. W. (1978). "A class of shrinkage estimators." **Journal Royal Statistical Society**, B, 40, 47-49.

Kawai, N. and Okamoto, M. (1979). "A generalization of the ridge function theorem." **Math. Japonica** 24, No.2, 175-178. [Abstract 79t-123. **Institute of Mathematical Statistics Bulletin** 8, No. 4.]

Trenkler, G. and Trenkler, D. (1981). "Estimable functions and reduction of mean squared error." **Methods of Operations Research** 44, 225-234. Oelgeschlager, Gunn & Hain, Cambridge, Mass.



Chapter 05: Normal Theory Maximum Likelihood

Bob Obenchain, Ph.D.
softR$_X$ freeware
13212 Griffin Run
Carmel, Indiana 46033-8835

Chapter 5: NORMAL THEORY MAXIMUM LIKELIHOOD

Here in Chapter 5 we discuss Normal-theory methods of identifying which shrinkage regression coefficient estimates are most likely to have optimal mean-squared-error, i.e. which set of shrinkage factors are most likely to be the target values, δ_i^{MSE}, of equation { 4.6 } for $1 \leq i \leq R$. Of course, once you have located the extent of shrinkage most-likely-to-be-optimal, you still have the option of limiting shrinkage using the "(2/R)ths Rule-of-Thumb" (defined in Section §4.2) and thus maximizing (in some weaker sense) the likelihood of ending up with an estimator that has good multivariate mean-squared-error characteristics.

The first three sections of Chapter 5 explore methods for classical (fixed-coefficient) linear models. Section §5.1 gives a brief review of the relatively well-known unrestricted maximum likelihood theory under which least-squares regression coefficients are BLUE (Best Linear Unbiased Estimates.) Section §5.2 starts our exploration of methods for statistical inference concerning shrinkage by first defining the "likelihood" that any given amount of shrinkage yields optimal mean-squared-error and by then deriving its maximum, as in Obenchain(1975). In Section §5.3, we develop a closed-form expression for the shrinkage estimator most likely to attain minimal mean-squared-error within the 2-parameter generalized shrinkage family, as in Obenchain(1981).

The last two sections of Chapter 5 explore models with random (stochastic) coefficients. Section §5.4 reviews Henderson's BLUP (Best Linear Unbiased Prediction) theory for "mixed" models (containing both fixed and random coefficients.) In practical applications, maximum likelihood estimates of mixed model coefficients usually end up being neither "linear" nor "unbiased". After all, just as in purely fixed-coefficient formulations, optimal estimates for random-coefficient models depend upon unknown parameters which, when estimated from the data at hand, yield operational analogs that are non-linear and biased. Section §5.5 develops much more detailed results on maximum likelihood estimation for the special case of purely-random models with a single variance component, as in Golub, Heath, and Wahba(1979) and Shumway(1982).

5.1 Unrestricted Maximum Likelihood and BLUE Theory

As we saw in Chapter 2, the general linear model is commonly stated using the pair of vector/matrix equations: $E(y \mid X) = 1\mu + X\beta$ and $V(y \mid X) = \sigma^2 I$. These are equations { 2.1 } and { 2.2 }, i.e. before the response vector, y, and the regressor matrix, X, are "centered" by subtracting off column means. As before, β is the column vector of unknown regression coefficients, μ is the unknown y intercept (the expected response corresponding to a

null row of X), and σ^2 is the unknown residual variance. Under Normal-distribution-theory, the joint likelihood function for μ, β and σ^2 is then

$$L(\mu, \beta, \sigma^2) = (2\pi\sigma^2)^{-N/2} e^{-u^2/2\sigma^2}, \qquad \{ 5.1 \}$$

where u^2 is the quadratic form

$$u^2 = (y - 1\mu - X\beta)^T (y - 1\mu - X\beta). \qquad \{ 5.2 \}$$

Also as before, we will use $\overline{y}$ and $\overline{x}^T$ to represent be the mean values of the original response values and regressor combinations, respectively. This allows us to decompose u^2 into a sum of two terms, thereby isolating the contributions to u^2 resulting from coordinates parallel to the 1 vector from those contributions from coordinates orthogonal to the 1 vector. Returning now to the convention that the y vector and the X matrix have been "centered", as in Section §2.1, equations { 2.3 } and { 2.4 }, we can rewrite equation { 5.2 } as

$$u^2 = (\overline{y} - \mu - \overline{x}^T\beta)^2 + (y - X\beta)^T (y - X\beta). \qquad \{ 5.3 \}$$

For any estimate $\widehat{\beta}$ of β, the first term in the above expression for u^2 can always be made to vanish simply by taking $\widehat{\mu} = \overline{y} - \overline{x}^T\widehat{\beta}$. As a result, we can drop μ from further, explicit consideration and view the likelihood as really being a function of only the β and σ^2 estimates.

It is well known that the global, unrestricted maximum of $L(\beta, \sigma^2)$ is then achieved at

$$L(b^o, \widehat{\sigma}^2) = (2\pi e\widehat{\sigma}^2)^{-N/2} \qquad \{ 5.4 \}$$

where $b^o = X^+y = G\,c$ is the least squares estimator of β [as in equation { 2.6 }] and $N \cdot \widehat{\sigma}^2 = y^Ty \cdot (1\text{-}R^2)$ is the minimum value for u^2 of equation { 5.3 }. In other words, the vector of ordinary least squares coefficients, b^o, is the maximum-likelihood estimator of the unknown, true β vector under Normal-distribution theory. Note that the maximum-likelihood estimator of σ^2, namely $\widehat{\sigma}^2 = y^Ty \cdot (1\text{-}R^2) / N$, is larger than the usual unbiased estimator [s^2 of equation { 2.22 }] by a multiplicative factor of $[N / (N - R - 1)]$. Although these results are quite well know, we will now review a derivation of { 5.4 } to illustrate some of the techniques we will also use in Section §5.2.

Our initial step toward maximizing the Normal-theory likelihood of equation { 5.1 } will be to minimize $u^2 = (y - X\widehat{\beta})^T (y - X\widehat{\beta})$ by choice of $\widehat{\beta}$ [and then set $\widehat{\mu} = \overline{y} - \overline{x}^T\widehat{\beta}$.] Thus note that the vector of partial derivatives of u^2 with respect to β is

$$\partial(u^2)/\partial\beta = 2 \cdot X^TX\beta - 2 \cdot X^Ty, \qquad \{ 5.5 \}$$

while the matrix of second partial derivatives is non-negative definite

$$\partial^2(u^2)/\partial\beta^2 = 2 \cdot X^TX. \qquad \{ 5.6 \}$$

Thus any solution to $\partial(u^2)/\partial\widehat{\beta} = 0$ yields the minimum value for u^2, and $\widehat{\beta} = b^o = X^+y$ is always a solution to these so-called "normal equations," $X^TX\,\widehat{\beta} = X^Ty$.

Our final step in maximizing the Normal-theory likelihood of equation { 5.1 } is to choose the estimator of σ^2. Equivalently, we can minimize the minus-twice-log-likelihood, $-2\cdot\ln(L) = N\cdot\ln(2\pi\sigma^2) + u^2/\sigma^2$. The first partial derivative is thus

$$\partial[-2\cdot\ln(L)]/\partial(\widehat{\sigma}^2) = \frac{N}{\widehat{\sigma}^2} - \frac{u^2}{(\widehat{\sigma}^2)^2}\,, \qquad \{\ 5.7\ \}$$

while the second partial derivative is

$$\partial^2[-2\cdot\ln(L)]/\partial(\widehat{\sigma}^2)^2 = -1\cdot\frac{N}{(\widehat{\sigma}^2)^2} + 2\cdot\frac{u^2}{(\widehat{\sigma}^2)^3}\,. \qquad \{\ 5.8\ \}$$

Note that $\partial[-2\cdot\ln(L)]/\partial(\widehat{\sigma}^2) = 0$ admits two solutions, $\widehat{\sigma}^2 = u^2/N$ and $\widehat{\sigma}^2 = +\infty$. But the infinite solution can be eliminated from consideration because the second partial derivative vanishes at this point. On the other hand, $\partial^2[-2\cdot\ln(L)]/\partial(\widehat{\sigma}^2)^2$ assumes the strictly positive value $[\,N/\widehat{\sigma}^4\,]$ at $\widehat{\sigma}^2 = u^2/N$ under the (almost certain) condition that the minimum sum-of-squares of residuals is strictly positive. This establishes that the maximum value for the likelihood of equation { 5.1 } is indeed given by { 5.4 }, where $\widehat{\mu} = \overline{y} - \overline{x}^Tb^o$, $b^o = X^+y$, and $\widehat{\sigma}^2 = y^Ty\cdot(1\text{-}R^2)/N$.

The least-squares estimator, b^o, is an unbiased estimator of β under the usual assumption that the given expectation equation, $E(y\,|\,X) = 1\,\mu + X\,\beta$, is correct. Similarly, under the usual assumption that the dispersion equation, $V(y\,|\,X) = \sigma^2\,I$, is correct, b^o is also the minimum variance estimator within the class of linear, unbiased estimators of β . As a result, b^o is commonly said to be the BLUE (Best Linear Unbiased Estimator) of β . Our derivation of equations { 5.4 } through { 5.7 } has actually established only that b^o is the (unrestricted) Normal-theory maximum likelihood estimator of β ; see Section §4a.2 of Rao(1973), pages 222-224, for the arguments that actually establish that all linear combinations, P^Tb^o, coincide with the BLUE of the corresponding estimable linear combinations, $P^T\beta$.

5.2 The Likelihood of Mean Squared Error Optimality

Under the assumption that our "target values" for optimal shrinkage are given by the $\delta_i^{MSE} = \gamma_i^2/(\gamma_i^2+\sigma^2\lambda_i^{-1})$ factors of equation { 4.6 }, we now wish to explore methods for identifying numerical values of shrinkage-regression-factors that are most likely to be ON TARGET under Normal-theory. Rather than simply paraphrase the arguments given in my papers, Obenchain(1975,1981), I use a new approach here that I hope will be somewhat easier to follow. To minimize potential for misinterpretation, let me describe the general point-of-view assumed below:

(i) Unrestricted maximum likelihood estimation of the parameters of a multiple regression model "uses up" a total of R+2 degrees-of-freedom. The estimate $b^o = X^+ y = G\,c$ (of the coefficient vector β) corresponds to R degrees-of-freedom. The estimates $\widehat{\mu} = \overline{y} - \overline{x}^T b^o$ (of the y intercept μ) and $\widehat{\sigma}^2 = y^T y \cdot (1\text{-}R^2) / N$ (of the residual variance σ^2) use 1 degree-of-freedom each.

(ii) Shrinkage regression estimation as viewed here always ends up "using" AT LEAST these same R+2 degrees-of-freedom. After all, shrinkage estimators are of the general form $b^\bigstar = G\,\Delta\,c$, where the c vector (containing the least-squares estimates of the true uncorrelated components γ of β) again corresponds to R degrees-of-freedom. Similarly, $\mu^\bigstar = \overline{y} - \overline{x}^T b^\bigstar$ uses up 1 degree-of-freedom. The residual sum-of-squares for shrinkage estimates, $(y - X b^\bigstar)^T (y - X b^\bigstar)$, can only exceed the minimum value, $y^T y \cdot (1\text{-}R^2)$, attained by unrestricted maximum likelihood. This excess lack-of-fit is of no real help in estimating the residual variance σ^2, so the shrinkage regression estimate of σ^2 "usually" defaults back to the least-squares estimate, $\widehat{\sigma}^2 = y^T y \cdot (1\text{-}R^2) / N$, with 1 degree-of-freedom.

(iii) Maximum likelihood methods for identifying minimal mean-squared-error shrinkage may "use up" as few as only 1 or 2 more degrees-of-freedom than the minimum number, R+2, listed above. Specifically, the elements of the diagonal matrix, Δ, of shrinkage factors employed in $b^\bigstar = G\,\Delta\,c$ may be functions of only 1 or 2 parameters. (For example, in Section §5.3, these two parameters will be Q and k; Q determines the shape/curvature of the shrinkage path, and k determines the extent of shrinkage along that path.) In any case, practical applications of shrinkage regression should always be thought of as using up a total of R+3, or R+4, or perhaps even more degrees-of-freedom. After all, they employ (restricted) estimates of Δ and μ as well as unrestricted estimates of the implied γ and σ^2.

(iv) The minimal mean-squared-error target value for shrinkage, Δ^{MSE}, is a nonlinear function of γ and σ^2. As a result, maximum likelihood search over the HIGHLY RESTRICTED parameter space defining shrinkage factors may "use up" as few as only 3 or 4 degrees-of-freedom. Three or four degrees-of-freedom is frequently much less than the MINIMUM of R+2 employed in an unrestricted maximum likelihood search. The estimates of γ and σ^2, denoted here by γ^{**} and σ^{**2}, that are derived within this restricted search are usually of very little interest in their own right. Instead, they are mere precursors that help us identify not only the restricted Δ factor values most likely to be Δ^{MSE} but also the minimum value for the corresponding minus-twice-log-likelihood-ratio (restricted likelihood divided by unrestricted likelihood.)

The Normal-theory likelihood for the uncorrelated components vector $\gamma = G^T \beta$ and the residual variance σ^2 [evaluated, again, at $\mu = \overline{y} - \overline{x}^T G\,\gamma$] is

$$L(\gamma, \sigma^2) = (2\pi\sigma^2)^{-N/2}\, e^{-u^2/2\sigma^2}, \qquad \{ 5.9 \}$$

where u^2 is the quadratic form

$$u^2 \quad = \; (y \; - \; H \, \Lambda^{1/2} \, \gamma)^T (y \; - \; H \, \Lambda^{1/2} \, \gamma) \, . \qquad \{ 5.10 \}$$

Now note that $\delta_i^{MSE} \; = \; \gamma_i^2 / (\gamma_i^2 + \sigma^2 \lambda_i^{-1})$ can be rewritten as

$$\gamma_i^2 \cdot \lambda_i \; = \; \sigma^2 \cdot [\delta_i^{MSE} / (1 \; - \; \delta_i^{MSE})] \qquad \{ 5.11 \}$$

or, equivalently, as

$$\gamma_i \; = \; \pm 1 \cdot \sigma \cdot \sqrt{\delta_i^{MSE} / [\lambda_i \cdot (1 \; - \; \delta_i^{MSE})]} \; . \qquad \{ 5.12 \}$$

In other words, knowing the numerical values of δ_1^{MSE}, δ_2^{MSE}, ..., δ_R^{MSE} would be tantamount to knowing the RELATIVE MAGNITUDES of the ABSOLUTE VALUES of the uncorrelated components, $|\gamma_1|$, $|\gamma_2|$, ..., $|\gamma_R|$, of β .

Equations { 5.11 } and/or { 5.12 } suggest the following FUNDAMENTAL DEFINITION for the likelihood that any given set of numerical values for shrinkage factors, δ_1, δ_2, ..., δ_R, coincide with the optimal mean-squared-error target values δ_1^{MSE}, δ_2^{MSE}, ..., δ_R^{MSE}. This likelihood is defined to equal the likelihood that

$$\gamma_i^{**} \; = \; \pm 1 \cdot \sigma^{**} \cdot \sqrt{\delta_i / [\lambda_i \cdot (1 \; - \; \delta_i)]} \, , \qquad \{ 5.13 \}$$

where this likelihood has been maximized by choice of the R numerical signs (positive or negative) of the corresponding uncorrelated component estimates and by choice of estimate, σ^{**} , for the residual standard deviation. As explained above, these γ_i^{**} and σ^{**2} estimates usually are of little interest themselves, at least when the δ_i factors have been restricted to lie within a 1- or 2-parameter shrinkage family. But { 5.13 } is the general expression that would apply even if the δ_i were totally unrestricted; see subsection §5.2.1 below for a description of the "cubic" estimator that results in this totally unrestricted case.

Note that relationship { 5.13 } allows us to rewrite the quadratic form, u^2 of equation { 5.10 }, as

$$u^2 \; = \; (y \; - \; H \, S \, \xi \, \sigma)^T (y \; - \; H \, S \, \xi \, \sigma) \, , \qquad \{ 5.14 \}$$

where S is a diagonal matrix of signs (i.e. diagonal elements of ± 1) and ξ is the vector of values defined by the following positive square-roots:

$$\xi_i = \sqrt{\delta_i / (1 \; - \; \delta_i)} \, , \qquad \{ 5.15 \}$$

for $1 \leq i \leq R$. Note, specifically, that the resulting maximum likelihood estimator of γ is being restricted to be of the general form

$$\gamma^{**} = \sigma^{**} \cdot S \, \Lambda^{-1/2} \, \xi \qquad \{ 5.16 \}$$

whenever the shrinkage factors that define ξ yield minimum mean-squared-error. Before continuing, let us note that relationships { 5.15 } and { 5.16 } are well defined only when all δ_i factors are strictly less than 1; after all, $\delta_i = 1$ would imply $\xi_i = +\infty$. In other words, the possibility "no-shrinkage-at-all" along any principal axis is automatically being excluded from consideration as a potential mean-squared-error-optimal extent for shrinkage. On the other hand, values that are (numerically) very close to 1 are not being rejected out of hand. Thus, whenever a relatively large shrinkage factor value like $\delta_i = 0.95$ or 0.99 is found to be "most likely" in the sense defined below, there may be no real difference of any practical (numerical) importance between the corresponding least-squares and optimally-shrunken estimators.

Noting that $H^T H = I$, $S^2 = I$, and $y^T H = \sqrt{y^T y} \cdot r^T$ as in { 2.15 } and { 2.16 }, we now rewrite u^2 of { 5.10 } and { 5.14 } again as

$$u^2 \;=\; y^T y \;-\; 2 \cdot \sqrt{y^T y} \cdot r^T S\, \xi\, \sigma^{**} + \sigma^{**2} \xi^T \xi \,. \qquad \{ 5.17 \}$$

To minimize u^2, the R numerical signs in S should be chosen to make the middle, negative term as large as possible...assuming, of course, that the σ^{**} estimate is strictly positive. Clearly, the $r^T S\, \xi = \sum r_{yi} \cdot s_i \cdot \sqrt{\delta_i / (1 - \delta_i)}$ factor is maximized by choice of these signs when $s_i = \text{sign}(r_{yi})$, yielding $r^T S\, \xi = \sum |r_{yi}| \cdot \sqrt{\delta_i / (1 - \delta_i)}$.

Finding the optimal, strictly positive estimate, σ^{**}, of the residual standard deviation is the final, remaining step in minimizing the restricted minus-twice-log-likelihood. The corresponding partial derivatives are

$$\partial[-2 \cdot \ln(L^{**})]/\partial\sigma^{**} \;=\; \frac{2 \cdot N}{\sigma^{**}} \;-\; \frac{2 \cdot y^T y}{\sigma^{**3}} \;+\; \frac{2 \cdot \sqrt{y^T y} \cdot \sum |r_{yi}| \cdot \xi_i}{\sigma^{**2}} \,, \qquad \{ 5.18 \}$$

and

$$\partial^2[-2 \cdot \ln(L^{**})]/\partial\sigma^{**2} \;=\; -\frac{2 \cdot N}{\sigma^{**2}} + \frac{6 \cdot y^T y}{\sigma^{**4}} \;-\; \frac{4 \cdot \sqrt{y^T y} \cdot \sum |r_{yi}| \cdot \xi_i}{\sigma^{**3}} \,. \qquad \{ 5.19 \}$$

Equating the first derivative to zero yields the three solutions $\sigma^{**} = +\infty$ and

$$\sigma^{**} \;=\; \sqrt{y^T y} \cdot \frac{-\sum |r_{yi}| \cdot \xi_i \pm \sqrt{(\sum |r_{yi}| \cdot \xi_i)^2 + 4 \cdot N}}{2 \cdot N} \,. \qquad \{ 5.20 \}$$

The negative solution is of no interest; this choice would make the middle term of { 5.17 } positive. And the infinite solution corresponds to an indeterminant second derivative of zero. The second derivative is strictly positive at the positive solution of { 5.20 } and, thus, minimum minus-twice-log-likelihood is achieved there, yielding:

$$\sigma^{**} \;=\; \frac{\sqrt{y^T y}}{2 \cdot N} \cdot \left[\sqrt{(\sum |r_{yi}| \cdot \xi_i)^2 + 4 \cdot N} \;-\; \sum |r_{yi}| \cdot \xi_i \right]$$

$$= \frac{2 \cdot \sqrt{y^T y}}{\left[\sqrt{(\sum |r_{yi}| \cdot \xi_i)^2 + 4 \cdot N} + \sum |r_{yi}| \cdot \xi_i \right]} . \qquad \{ 5.21 \}$$

Of course, the resulting restricted minimum of the likelihood in { 5.9 } cannot be smaller than the unrestricted minimum, { 5.4 } . Therefore, the resulting (non-negative) minus-twice-log-likelihood-ratio statistic will be of the form

$$-2 \cdot \ln(L^{**} / \hat{L}) = N \cdot \ln(\sigma^{**2} / \hat{\sigma}^2) + \xi^T \xi - \sqrt{y^T y} \cdot \sum |r_{yi}| \cdot \xi_i / \sigma^{**}, \quad \{ 5.22 \}$$

and this statistic will have an asymptotic chi-squared distribution (as N increases to ∞) with degrees-of-freedom equal to R minus the number of "free" parameters remaining among the δ_1, δ_2, ..., δ_R factors under any restriction that might be imposed.

5.2.1 Unrestricted Maximum Likelihood Shrinkage: The Cubic Estimator

When no restrictions whatsoever are placed upon the δ-factors, { 5.22 } will be zero and will have zero degrees-of-freedom. To see this, note that the unrestricted maximum likelihood estimate of $\delta_i^{MSE} = \gamma_i^2/(\gamma_i^2+\sigma^2 \lambda_i^{-1})$ is clearly $c_i^2/(c_i^2+\hat{\sigma}^2 \lambda_i^{-1}) = r_{yi}^2 / [r_{yi}^2 + (1 - R^2)/N]$ where $c = G^T b^o$ is the vector of least-squares estimates for uncorrelated components and $\hat{\sigma}^2 = y^T y \cdot (1-R^2)/N$ is the least-squares residual mean square of { 5.4} . The corresponding values of the ξ_i terms are $\xi_i = |r_{yi}| \cdot \sqrt{N/(1 - R^2)}$. Therefore

$$\sum |r_{yi}| \cdot \xi_i = R^2 \cdot \sqrt{N/(1 - R^2)} \quad \text{and} \quad \sqrt{(\sum |r_{yi}| \cdot \xi_i)^2 + 4 \cdot N} = (2 - R^2) \cdot \sqrt{N/(1 - R^2)} .$$

In other words, the unrestricted estimates are $\sigma^{**2} = \hat{\sigma}^2$ in { 5.21 } and $\gamma^{**} = c$ in { 5.16 } .

The corresponding maximum-likelihood shrinkage estimator, $\hat{\delta}_i c_i$, for γ_i would be $c_i^3/(c_i^2+\hat{\sigma}^2 \lambda_i^{-1})$, which is a specific nonlinear estimator of "cubic" form. Thompson(1968), Figures 1 and 2 [pages 116 and 117], gave Normal-theory mean-squared-error plots (computed using numerical integration) for this special form of nonlinear estimator, where his horizontal axis was $|\gamma|/\sigma$ and his plotting range was unbounded (0 to ∞.) Dwivedi, Srivastava, and Hall(1980) and Hemmerle and Carey(1981) also study the mean-squared-error properties of this cubic estimator. See Figure XX, Chapter 6, for a plot of the simulated mean-squared-error-risk of the cubic estimator versus δ^{MSE}, i.e. over the finite range from 0 to 1.

5.2.2 Maximum Likelihood UNIFORM Shrinkage

Under the "uniform shrinkage" restriction that $\delta_1 = \delta_2 = ... = \delta_R$, this common shrinkage factor can be written as $\delta = 1 / (1 + k)$. As a result, $\xi_1 = \xi_2 = ... = \xi_R = \sqrt{\delta / (1 - \delta)} =$

$k^{-1/2}$, and the value of k that minimizes { 5.22 } [with chi-square degrees-of-freedom = R − 1 in the limit as n approaches ∞] is

$$k^{**} = (1 - R \cdot \overline{|r|}^2) / (N \cdot \overline{|r|}^2) , \qquad \{ 5.23 \}$$

where $\overline{|r|} = (|r_{y1}|+|r_{y2}|+...+|r_{yR}|) / R$ is the average of the absolute values of the principal correlations, Obenchain(1975). The corresponding minimum minus-twice-log-likelihood-ratio is then $-2 \cdot \ln(L^{**} / \widehat{L}) = N \cdot \ln[1 + \frac{(R-1)}{(N-R-1)} S]$, where S is the "uniform shrinkage statistic" of Obenchain(1975) :

$$S = \frac{(N-R-1) \cdot \sum (|r_{yi}| - \overline{|r|})^2}{(R-1) \cdot (1 - R^2)} . \qquad \{ 5.24 \}$$

No derivation of equations { 5.23 } and { 5.24 } will be given now because they are simple special cases [Q=1] of the general theory derived below in Section §5.3.

The corresponding restricted estimates of σ and of the uncorrelated components of β would be

$$\sigma^{**2} = y^T y \cdot (1 - R \cdot \overline{|r|}^2) / N \qquad \{ 5.25 \}$$

and

$$\gamma_i^{**} = \sqrt{y^T y / \lambda_i} \cdot \text{sign}(r_{yi}) \cdot \overline{|r|} \quad ...\text{for } 1 \leq i \leq R. \qquad \{ 5.26 \}$$

Again, these latter restricted estimates (the variance and the components) are of little real interest. After all, the numerical value of the maximum-likelihood "common" shrinkage factor, $\delta^{**} = N \cdot \overline{|r|}^2 / [1+(N - R) \cdot \overline{|r|}^2]$, would actually be used in conjunction with the UNRESTRICTED maximum-likelihood component estimates (c_1 , c_2 , ..., c_R) and the UNRESTRICTED residual variance, $\widehat{\sigma}^2$, of { 5.4 }. In other words, the restricted σ^{**2} and γ_i^{**} estimates are of interest ONLY in the sense that they help to define the minimum minus-twice-log-likelihood-ratio via equation { 5.24 } .

At the time of my original publication on Normal-theory maximum likelihood methods for shrinkage regression, Obenchain(1975), the only known closed-form solutions to the general expressions { 5.16 } and { 5.21 } were the two special cases treated above in subsections §5.2.1 and §5.2.2. Thus, I proposed a general technique I called <u>likelihood monitoring</u> for applying equations { 5.16 } and { 5.21 } to any parametric family of generalized shrinkage factors. This approach simply involves actual numerical computation of { 5.16 }, { 5.21 } and { 5.22 } upon a lattice of numerical values for δ_1 , δ_2 , ..., δ_R . These sorts of computations can be tedious, of course, but the vast majority of criteria that have been proposed for choosing a "best" shrinkage estimator are commonly applied in this computationally-intensive fashion.

5.3 Closed Form Expressions within the 2-Parameter Family

Within the 2-parameter family of Goldstein and Smith(1974), shrinkage factors are of the general form $\delta_i = 1 / (1 + k \cdot \lambda_i^{Q-1})$ of equation { 3.8 }, where the power, Q, determines the SHAPE/CURVATURE of the shrinkage path through likelihood space while the k factor determines the EXTENT of shrinkage. The natural range for the k parameter is all the way from k=0 for "no shrinkage" to the limit as k approaches $+\infty$, where all δ_i factors are shrunken "completely" to 0. The special case of shape Q=1 for "uniform shrinkage" was described in subsection §5.2.2. Among the other common choices for Q described in Section §3.3 of Chapter 3 are Q=0 for "ordinary ridge regression," Hoerl and Kennard(1970), and the limit as Q approaches $-\infty$ for "principal components regression," Marquardt(1970). In practical applications to ill-conditioned regression problems where the eigenvalue spectrum of the regressor X^TX matrix is wide, values of Q outside of the range $-5 \leq Q \leq +5$ rarely need to be considered. At the other extreme, where absolutely no ill-conditioning is present because $\lambda_1=\lambda_2=...=\lambda_R$, all values of Q would yield the same, uniform shrinkage pattern.

5.3.1 The most-likely-to-be-mse-optimal shrinkage extent, k, for given shape/curvature.

When $\delta_i = 1 / (1 + k \cdot \lambda_i^{Q-1})$, the ξ_i terms of { 5.15 } are of the general form $\xi_i = \sqrt{\delta_i / (1 - \delta_i)} = \sqrt{\lambda_i^{(1-Q)} / k}$. Therefore { 5.16 } becomes

$$\gamma_i^{**} = \pm \sigma^{**} / \sqrt{k \cdot \lambda_i^{Q/2}} . \qquad \{ 5.27 \}$$

There is a redundancy in { 5.27 } between k and the estimate of σ that could not be fully exploited in deriving a general expression like that given by equation { 5.21).

Let us now denote the common, unknown value of $\gamma_i^2 \lambda_i^Q = \sigma^{**2}/k$ in { 5.27 } by ρ_Q^2 :

$$\rho_Q^2 = \gamma_1^2 \lambda_1^Q = \gamma_2^2 \lambda_2^Q = ... = \gamma_R^2 \lambda_R^Q . \qquad \{ 5.28 \}$$

Equation { 5.27 } can then be rewritten as $\gamma_i^{**} = \pm \rho_Q^{**} \cdot \lambda_i^{-Q/2}$ and, again using $s_i = \text{sign}(r_{yi})$, the general residual-sum-of-squares equation of { 5.10 }, { 5.14 } and { 5.17 } now becomes

$$u^2 = y^Ty - 2 \cdot \sqrt{y^Ty} \cdot \rho_Q^{**} \cdot \sum |r_{yi}| \cdot \lambda_i^{(1-Q)/2} + \rho_Q^{**2} \cdot \sum \lambda_i^{(1-Q)} . \qquad \{ 5.29 \}$$

The corresponding partial derivatives of u^2 are

$$\partial [u^2] / \partial \rho_Q^{**} = -2 \cdot \sqrt{y^Ty} \cdot \sum |r_{yi}| \cdot \lambda_i^{(1-Q)/2} + 2 \cdot \rho_Q^{**} \cdot \sum \lambda_i^{(1-Q)} , \qquad \{ 5.30 \}$$

and

$$\partial^2 [u^2] / \partial \rho_Q^{**2} = +2 \cdot \sum \lambda_i^{(1-Q)} . \qquad \{ 5.31 \}$$

Thus the minimum value of u^2 clearly occurs at $\partial[u^2]/\partial\rho_Q^{**} = 0$, which implies

$$\rho_Q^{**} = \sqrt{y^T y} \cdot \frac{\sum |r_{yi}| \cdot \lambda_i^{(1-Q)/2}}{\sum \lambda_i^{(1-Q)}} \ . \qquad \{ 5.32 \}$$

and

$$u^{**2} = \text{minimum}(u^2) = y^T y \cdot [1 - R^2 \cdot CRL^2(Q)] , \qquad \{ 5.33 \}$$

where CRL(Q) is the "curlicue" function that measures CORRELATION between the vector of absolute values of the principal correlations and the vector of regressor singular values raised to the (1 − Q)/2-th power. Specifically,

$$CRL(Q) = \frac{\sum |r_{yi}| \cdot \lambda_i^{(1-Q)/2}}{\sqrt{\sum r_{yi}^2 \cdot \sum \lambda_i^{(1-Q)}}} \ . \qquad \{ 5.34 \}$$

Note that this correlation can also be viewed as the Cosine of the angle between the "R-vector" of absolute principal axis correlations and the "L-vector" of regressor eigenvalues raised to a power determined by the path shape/curvature parameter, Q ; this notation motivates the CRL(Q) mnemonic, Obenchain(1981). [In Obenchain(1975), equation (4.6), CRL(Q) was denoted by COS(q).]

Our final step in maximizing the restricted Normal-theory likelihood is to choose the estimator of σ^{**2} given the minimum u^2, which is the exact same sort of problem we treated in equations { 5.7 } and { 5.8 }. Of the two possible solutions to $\partial[-2 \cdot \ln(L^{**})]/\partial(\sigma^{**2}) = 0$, the $\sigma^{**2} = +\infty$ solution is again ruled out in favor of $\sigma^{**2} = u^{**2}/ N$. As a result, the most-likely-to-be-optimal extent of shrinkage along the path of shape Q is given by:

$$k^{**} = \sigma^{**2}/\rho_Q^{**2} = [\sum \lambda_i^{(1-Q)}] \cdot \frac{[1 - R^2 \cdot CRL^2(Q)]}{[N \cdot R^2 \cdot CRL^2(Q)]} \ . \qquad \{ 5.35 \}$$

The corresponding minimum minus-twice-log-likelihood-ratio is then $-2 \cdot \ln(L^{**} / \widehat{L}) = N \cdot \ln[1 + \frac{(R-1)}{(N-R-1)} S(Q)]$, where S(Q) becomes :

$$S(Q) = \frac{(N-R-1) \cdot R^2 [1 - CRL^2(Q)]}{(R-1) \cdot (1 - R^2)} \ . \qquad \{ 5.36 \}$$

Note that { 5.36 } agrees with equation (4.6) of Obenchain(1975), but the closed-form expression, { 5.35 }, was unknown at that time. This most-likely-to-be-mse-optimal value of k given Q was first derived in Obenchain(1981), equation (2.5).

5.3.2 The most-likely-to-be-mse-optimal shrinkage shape/curvature, Q.

It is clear from { 5.36 } that the minimum minus-twice-log-likelihood-ratio depends upon Q only through the curlicue function of { 5.34 } :

$$CRL(Q) = \frac{\sum |r_{yi}| \cdot \lambda_i^{(1-Q)/2}}{\sqrt{\sum r_{yi}^2 \cdot \sum \lambda_i^{(1-Q)}}} \quad .$$

And it is clear from this definition that CRL(Q) cannot be made negative by choice of Q. As a result, S(Q) is minimized (as is u^{**2} of { 5.33 }) by choice of Q by making CRL(Q) as large as possible.

Before going further, perhaps we should discuss why choosing the Q-shape so as to maximize the curlicue function represents intuitive "common sense." Notice, first, that the ordered regressor eigenvalues, $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_R > 0$, will have very little influence upon CRL(Q) whenever Q is close to +1 ...because the $\lambda_i^{(1-Q)/2}$ values are all equal to 1 at Q=+1. Whenever the absolute values of all of the principal correlations are "nearly" equal, the angle between the |R|-vector and the Q=+1 L-vector $\propto 1$ will be small and, thus, CRL(Q) will be maximized at a Q value close to +1. And Q=+1 represents uniform shrinkage, $\delta_1=\delta_2=...=\delta_R$.

Next, note that the shrinkage factors will be monotonically non-increasing, $\delta_1 \geq \delta_2 \geq ... \geq \delta_R$, when Q is strictly less than +1 and monotonically non-decreasing, $\delta_1 \leq \delta_2 \leq ... \leq \delta_R$, when Q is strictly greater than +1. After all, these factors are being restricted to be of the functional form $\delta_i = 1 / [\, 1 + k \cdot \lambda_i^{(1-Q)} \,]$.

> Therefore, when the "trailing" absolute principal correlations [$|r_{yR}|$, $|r_{y(R-1)}|$, ...] are relatively large, CRL(Q) will tend to be maximized at values of Q greater than +1. These are the somewhat pathological cases where the regressor coordinates that have the least adequate spread in their numerical values are the coordinates most highly correlated with the response.
>
> But, when the "leading" absolute principal correlations [$|r_{y1}|$, $|r_{y2}|$, ...] are relatively large, CRL(Q) will tend to be maximized at values of Q less than +1 . In fact, the Q that maximizes CRL(Q) may be less than 0 in these cases. These are the "business-as-usual" cases where the regressor coordinates that have the most adequate spread in their numerical values are most highly correlated with the response. After all, this is the strategy you would use in a "designed experiment" ...where you would deliberately explore responses over a relatively wide range for any/all relatively important "factors."

In actual applications of normal-theory maximum-likelihood to ill-conditioned regression problems, I favor considering only a limited number of possible shapes, Q . For example, my personal computer application, RXridge, considers only integer and half-integer values within the range $-5 \leq Q \leq +5$. I see no practical reason for ever considering a finer lattice of Q shapes than, say, 0.1 (one place after the decimal) or a wider range of Q shapes than $-5 \leq Q \leq +5$; but this is, perhaps, mostly a matter of personal taste. Anyway, whenever a

limited number of Q shapes are under consideration, the most straight-forward way to find the corresponding restricted maximum of CRL(Q) is simply to compute all of these values ...then pick the Q shape yielding the largest CRL(Q) value.

For those shrinkage regression practitioners who simply cannot resist the temptation to locate "the" optimal Q shape (with great numerical precision), a Newtonian descent method for iterative search can be used. Specifically, with Q_s^{**} denoting the best estimate of the Q that maximizes CRL(Q) at stage s of the iteration, the update equation for step s+1 becomes

$$Q_{s+1}^{**} \;=\; Q_s^{**} \;-\; [\, CRL'(Q) \,/\, CRL''(Q) \,] \,, \qquad \{ 5.37 \}$$

where

$$CRL'(Q) = \partial CRL(Q)/\partial Q = \sum_{i=1}^{R} \left[\frac{|r_{yi}| \cdot \lambda_i^{(1-Q)/2}}{R \cdot \sqrt{\sum \lambda_i^{(1-Q)}}} \cdot (H_i + \tfrac{J}{2}) \right] , \qquad \{ 5.38 \}$$

for

$$R^2 = \sum r_{yj}^2 \,, \quad H_i = \frac{-\ln(\lambda_i)}{2} \,, \quad J = \frac{\sum \lambda_j^{(1-Q)} \cdot \ln(\lambda_j)}{\sum \lambda_j^{(1-Q)}} \,,$$

and

$$CRL''(Q) = \partial^2 CRL(Q)/\partial Q^2 = \sum_{i=1}^{R} \left[\frac{|r_{yi}| \cdot \lambda_i^{(1-Q)/2}}{R \cdot \sqrt{\sum \lambda_i^{(1-Q)}}} \cdot \left\{ (H_i + \tfrac{J}{2})^2 + \tfrac{J^2}{2} \;-\; K \right\} \right] , \qquad \{ 5.39 \}$$

for

$$K = \frac{\sum \lambda_j^{(1-Q)} \cdot \ln^2(\lambda_j)}{\sum \lambda_j^{(1-Q)}} \,.$$

Convergence to a (possibly local) maximum of CRL(Q) requires finding a shape value Q^{**} such that $CRL'(Q^{**})=0$ and $CRL''(Q^{**}) < 0$. The step-size, $[CRL'(Q)/CRL''(Q)]$, in { 5.37 } should be bisected whenever $CRL(Q_{s+1}^{**})$ fails to achieve an increase over $CRL(Q_s^{**})$, and the search direction should be reversed when $CRL''(Q_s^{**}) > 0$.

5.3.3 The limit as the shrinkage shape/curvature, Q, approaches $-\infty$.

Note that CRL(Q) can also be thought of as the cosine of the angle between the vector of absolute principal correlations, [$|r_{y1}|$, $|r_{y2}|$, ..., $|r_{yR}|$], and the following vector of powers of eigenvalue ratios, [1, $(\lambda_2/\lambda_1)^{(1-Q)/2}$, $(\lambda_3/\lambda_1)^{(1-Q)/2}$, ..., $(\lambda_R/\lambda_1)^{(1-Q)/2}$]. This latter vector clearly approaches [1, 0, 0, ..., 0] as Q approaches $-\infty$ whenever $\lambda_1 > \lambda_2$, i.e. when the leading eigenvalue of the centered-regressor X^TX matrix is larger than all of the other eigenvalues. Thus CRL(Q) approaches $|r_{y1}|/R$ as Q approaches $-\infty$; similarly, $k^{**} \cdot \lambda_1^{(Q-1)}$

approaches [1 $- r_{y1}^2$]/[N $\cdot r_{y1}^2$] while $k^{**} \cdot \lambda_j^{(Q-1)}$ approaches $+\infty$ for j = 2, 3, ..., R. The shrinkage factors most-likely-to-be-mse-optimal in this limit are thus $\delta_1^{**}(-\infty) = N / [N - 1 + r_{y1}^{-2}]$ and $\delta_2^{**}(-\infty) = \delta_3^{**}(-\infty) = ... = \delta_R^{**}(-\infty) = 0$, which is a point on the principal components regression path, Massy(1965), that has a Marquardt(1970) fractional rank of less than 1 .

5.3.4 Large Sample Chi-Squared Tests of MSE-Optimality

A large sample χ^2 (Chi-Squared) test can be based upon the minimum value of the minus-twice-log-likelihood-ratio, $-2 \cdot \ln(L^{**} / \widehat{L}) = N \cdot \ln[1 + \frac{(R-1)}{(N-R-1)} S(Q)]$, where S(Q) is defined as in { 5.36 } . The degrees-of-freedom used in this test would be (R – 1) if one's shrinkage shape parameter, Q, had been selected without reference to the observed response data. But the appropriate degrees-of-freedom would be (R – 2) if, instead, S(Q) has been minimized by choice of Q. Whenever this χ^2 statistic is significantly greater than zero, statistical evidence has been accumulated suggesting that the 2-parameter family is "too restrictive" to contain the MSE-optimal values for the shrinkage factors.

5.4 Maximum Likelihood Methods for Mixed Linear Models

Statistical literature on the subject of mixed linear models (i.e. models containing both fixed and random coefficients) has been growing for 40-50 years; several major, new contributions to this area have appeared within the last 25 years. Henderson(1950) introduced the mixed model equations; his more recent fundamental contributions, Henderson(1975, 1984, 1990), include BLUP theory. Rao(1971a,b) introduced MINQUE and MIVQUE estimates of variance components; Patterson and Thompson(1971) defined REML estimation; Searle(1971, 1979, 1988) unified the theory of mixed, linear models and variance components; Harville(1977, 1988, 1990) has provided maximum likelihood theory and algorithms as well as prediction methodology; and Robinson(1990) has provided a highly readable BLUP review article. The vast majority of technical details on normal-distribution-theory maximum-likelihood estimation for mixed linear models will be postponed until Chapter 7 rather than being presented here in Chapter 5. However, we will introduce sufficient material, here in Chapter 5, to establish a few key parallels between the otherwise distinct maximum-likelihood approaches to fixed coefficient and random coefficient models.

A mixed linear model can be written in the general form:

$$y = X \cdot \beta + Z \cdot \theta + \eta \qquad \{ 5.40 \}$$

where

$$Var(y) = V = Z \cdot Var(\theta) \cdot Z^T + Var(\eta) \qquad \{ 5.41 \}$$

and the variance matrices, $G = Var(\theta)$ and $R = Var(\eta)$, are positive-definite matrices (frequently of block-diagonal form) that containing known or unknown parameters, generally called "variance components."

Now, the "unified" theory of mixed linear models tells us that the BLUE's and BLUP's are solutions to the Henderson(1975, 1984) MIXED MODEL EQUATIONS

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{bmatrix} \cdot \begin{bmatrix} \hat{\beta} \\ \hat{\theta} \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{bmatrix} . \qquad \{ 5.42 \}$$

We can display closed form solutions to these equations under the assumption/pretense that the G and R matrices are known matrices. Namely, the BLUEs would be

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y , \qquad \{ 5.43 \}$$

$$= \left[X^T (R + Z G Z^T)^{-1} X \right]^{-1} X^T (R + Z G Z^T)^{-1} y ,$$

and the BLUPs would be

$$\hat{\theta} = G Z^T V^{-1} \left[y - X \hat{\beta} \right] , \qquad \{ 5.44 \}$$

$$= (Z^T R^{-1} Z + G^{-1})^{-1} \left[Z^T R^{-1} - Z^T R^{-1} X W X^T (R + Z G Z^T)^{-1} \right] y ,$$

where $V = R + Z G Z^T$ from { 5.41 } and $W = \{ X^T (R + Z G Z^T)^{-1} X \}^{-1}$.

In his recent review article, Robinson(1991) stresses that Henderson's BLUP terminology is "reasonable" in the sense that $\hat{\theta}$ of { 5.44 } is indeed a Linear and Unbiased estimator of the random θ vector of { 5.6 } that is Best in a minimum variance sense...but only when the G and R matrices are formed using the UNKNOWN, TRUE values for all variance components! [By the way, $\hat{\theta}$ is termed a Predictor (rather than an estimator) primarily because θ is random rather than a fixed effect (unknown constant.)]

In reality, the variance components are frequently not only unknown but also the primary focus of one's attention in both estimation and statistical inference! Because numerical estimates for variance components (derived from the data at hand) are inserted into { 5.43 } and { 5.44 }, in practical applications neither $\hat{\beta}$ nor $\hat{\theta}$ is actually linear, neither $\hat{\beta}$ nor $\hat{\theta}$ is actually unbiased, and neither $\hat{\beta}$ nor $\hat{\theta}$ is actually best in any minimum variance sense! In other words, BLUE and BLUP terminology is truly "unfortunate" when applied to mixed model estimation.

An interesting facet of mixed model estimation that is implied by (but, perhaps, not immediately obvious from) equations { 5.42 }, { 5.43 } and { 5.44 } is that BLUPs are forms of "shrinkage estimators." This analogy is, perhaps, most obvious in the following special case.

5.5 Completely Random Models with a Single Variance Component

Golub, Heath and Wahba(1979) display a (random coefficient) maximum likelihood criterion for picking the extent of shrinkage in ridge regression [their equation (5.3)] that I, at least, found quite mysterious until I studied Shumway(1982). The mixed model considered by these authors is "completely random" in the sense that the $X \cdot \beta$ term of { 5.40 } contains only the (rank 1) overall mean, $1 \cdot \mu$. Furthermore, the random θ coefficient variation involves a "single" (unknown) variance component, σ_θ^2. Thus Var(θ) $= G = \sigma_\theta^2 \cdot D^{-1}$ where D^{-1} represents the known dispersion structure of the unknown θ vector. In other words, all coefficients are assumed to have known intercorrelations and known relative variances. The special case of uncorrelated, homoscedastic coefficients is $D = I$.

In exactly the same way that the non-constant columns of X are usually "centered," we suppose now that the columns of Z have been made to sum to 0 by subtracting off column means. To avoid complications unnecessary to this discussion, suppose that the centered Z matrix if of full (column) rank, P_z. Finally, suppose that the η disturbance terms are uncorrelated and homoscedastic: $R = \sigma^2 \cdot I$, as in { 2.2 }, where σ^2 is the unknown error variance component.

5.5.1 Demonstration that BLUP estimates are shrinkage estimates in this case.

Under the above assumptions, $X^T R^{-1} Z = 0$ in { 5.42 }, so that the matrix equations for $\widehat{\mu}$ and $\widehat{\theta}$ "uncouple" as follows. The top equation in { 5.42 } reduces to $\widehat{\mu} = 1^T y / 1^T 1 = \overline{y}$, and the bottom P_z equations yield:

$$\widehat{\theta} = (Z^T Z + (\sigma^2/\sigma_\theta^2) \cdot D)^{-1} Z^T y . \qquad \{ 5.45 \}$$

The second matrix expression in { 5.44 } is equivalent to { 5.45 } because $Z^T R^{-1} X = 0$. Notice also that, because the Z matrix has been centered, replacing the response vector y by ($y - 1 \cdot \overline{y}$) in { 5.45 } would not change the $\widehat{\theta}$ estimate.

Now note in equation { 5.45 } that:

(i) the variance component ratio, $\sigma^2/\sigma_\theta^2 = \phi^{-2}$, plays the role of the "k" (shrinkage-extent) factor of { 3.9 } and { 5.27 }, while

(ii) D plays the role of the $(Z^T Z)^Q$ matrix in equation { 3.9 } for the “Q-shape" of equation { 5.27 }.

In particular, $D = I$ in { 5.45 } yields the random-coefficient version of the (ordinary) ridge regression shrinkage path [Q=0] of Hoerl and Kennard(1970). Therefore, we have established that BLUP estimates are shrinkage estimates ...at least in the case of random coefficient models with a single variance component. That more complicated forms of BLUP also represent shrinkage can be easily verified via numerical computation.

5.5.2 Random coefficient maximum likelihood choice of shrinkage extent.

The Normal-theory joint likelihood function for the responses, y , can be written in the form

$$L(\theta, \sigma_\theta^2, \sigma^2) = (2\pi|V|)^{-1/2} e^{-u^2/2}, \qquad \{ 5.46 \}$$

where u^2 is the quadratic form

$$u^2 = (y - 1\cdot\mu)^T V^{-1} (y - 1\cdot\mu), \qquad \{ 5.47 \}$$

and $V = \sigma_\theta^2 \cdot Z D^{-1} Z^T + \sigma^2 \cdot I$ from { 5.41 } .

Writing $k = \sigma^2/\sigma_\theta^2$ and using well-known determinant and matrix-inverse identities [Rao(1973), pages 32 and 33], it follows that

$$|D|\cdot|V| = \begin{vmatrix} \sigma^2 I & \sigma_\theta Z \\ -\sigma_\theta Z^T & D \end{vmatrix} = \sigma^{2\cdot N} \cdot |D + k^{-1}\cdot Z^T Z|, \qquad \{ 5.48 \}$$

and

$$V^{-1} = \sigma^{-2}\cdot I - \sigma^{-2}\cdot Z (Z^T Z + k\cdot D)^{-1} Z^T. \qquad \{ 5.49 \}$$

These expressions allow us to rewrite { 5.46 } as

$$-2\cdot\ln(L) = \ln(2\pi) + N\cdot\ln\sigma^2 - \ln|D| - P_z\cdot\ln k + \ln|Z^T Z + kD| + u^2/\sigma^2, \qquad \{ 5.50 \}$$

where $u^2 = \sum(y_j - \mu)^2 - y^T Z (Z^T Z + k\cdot D)^{-1} Z^T y$. This minus-twice-log-likelihood is minimized, first, by taking $\hat{\mu} = \overline{y}$ to minimize u^2 for any given value of k. Then, exactly as in equations { 5.7 } and { 5.8 }, the minimizing error variance component is $\hat{\sigma}^2 = \hat{\sigma}^2(k) = [\sum(y_j - \overline{y})^2 - y^T Z \hat{\theta}(k)] / N$ for $\hat{\theta}(k) = \hat{\theta}$ of { 5.45 }. Substituting these values into { 5.50 } yields an expression for the minus-twice-log-likelihood that is a function of k only :

$$-2\cdot\ln(L) = \ln(2\pi) + N - \ln|D| + N\cdot\ln\hat{\sigma}^2(k) - P_z\cdot\ln k + \ln|Z^T Z + k\cdot D|, \qquad \{ 5.51 \}$$

as in equation (18) of Shumway(1982). Numerical search over a lattice of alternative values for k would then be used to locate the (approximate) minimum of { 5.51 }.

Rather than base computations on this minus-two-log-likelihood expression, Golub, Heath and Wahba(1979) suggest minimizing the equivalent criterion:

$$M(k) = \frac{1}{N}\cdot \frac{(y-\overline{y}\cdot 1)^T (I - A(k)) (y-\overline{y}\cdot 1)}{|I - A(k)|^{1/N}}, \qquad \{ 5.52 \}$$

where A(k) = Z ($Z^T Z + k \cdot D$)$^{-1}$ Z^T . Note that the numerator of { 5.52 } is simply $N \cdot \hat{\sigma}^2(k)$ and that A(k) = I $- \sigma^2 V^{-1}$ by { 5.49 }. Thus the N-th root of the determinant can be rewritten as $| I - A(k) |^{1/N} = |\sigma^2 V^{-1}|^{1/N}$ and this, in turn, is equivalent, by { 5.48 }, to the product of terms [$|D|^{1/N} \cdot k^{P_Z/N}/|D \cdot k + Z^T Z|^{1/N}$] whose negative logarithm is contained in { 5.51 } when that expression is divided by N.

REFERENCES for Chapter Five

Golub, G.H., Heath, M., and Wahba, G. (1979). "Generalized cross-validation as a method for choosing a good ridge parameter." **Technometrics** 21, 215-223.

Henderson, C. R. (1950). "Estimation of genetic parameters (abstract.)" **Annals of Mathematical Statistics** 21, 309-310.

Henderson, C. R. (1973). "Sire evaluation and genetic trends." In **Proceedings of the Animal Breeding and Genetics Symposium in Honor of Dr. Jay L. Lush** 10-41. Amer. Soc. Animal Sci. – Amer. Dairy Sci. Assoc. – Poultry Sci. Assn., Champaign, Illinois.

Henderson, C. R. (1984). **Applications of Linear Models in Animal Breeding**, University of Guelph.

Henderson, C. R. (1990). "Statistical methods in animal improvement: historical overview." In **Advances in Statistical Methods for Genetic Improvement in Livestock.** Springer-Verlag 1-14, 413-436.

Obenchain, R. L. (1975). "Ridge analysis following a preliminary test of the shrunken hypothesis." **Technometrics**, 17, 431-441. (Discussion: McDonald, G. C., 443-445.)

Obenchain, R. L. (1981). "Maximum likelihood ridge regression and the shrinkage pattern hypotheses." Abstract 81t-23. **I.M.S. Bulletin** 10, 37.

Obenchain, R. L. (1984). "Maximum likelihood ridge displays." **Communications in Statistics A**, 13, 227-240. (Proceedings of the Fordham Ridge Symposium, ed. H. D. Vinod.)

Rao, C. R. (1973). **Linear Statistical Inference and its Applications,** 2nd edition. New York: John Wiley & Sons.

Searle, S. R. (1971). **Linear Models**. New York: John Wiley and Sons.

Shumway, R. H. (1982). "Maximum likelihood estimation of the ridge parameter in linear regression." **Technical Report, Department of Statistics**, University of California at Davis.

Further Reading for Chapter Five

Dwivedi, T. D., Srivastava, V. K. and Hall, R. L. (1980). "Finite sample properties of ridge estimators." **Technometrics** 22, 205-212.

Goldstein, M. and Smith, A. M. F. (1974). "Ridge-type estimators for regression analysis." **Journal Royal Statistical Society B**, 36, 284-291.

Fuller, W. A. and Battese, G. E. (1973). "Transformations for estimation of linear models with nested error structure," **Journal of the American Statistical Association**, 68, 626-632.

Harville, D. A. (1977). "Maximum likelihood approaches to variance component estimation and to related problems," **Journal of the American Statistical Association** 72, 320-338.

Harville, D. A. (1986). "Using least squares software to compute combined intra-interblock estimates of treatment contrasts," **The American Statistician**, 40, 153-157.

Hemmerle, W. J. and Carey, M. B. (1981). "Some properties of generalized ridge estimators." Department of Computer Science and Experimental Statistics, University of Rhode Island.

Hoerl, A. E. and Kennard, R. W. (1970a). "Ridge regression: biased estimation for nonorthogonal problems." **Technometrics** 12, 55-67.

Jennrich, R. I. and Schluchter, M. D. (1986). "Unbalanced repeated-measures models with structured covariance matrices," **Biometrics**, 42, 805-820.

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

Massy, W. F. (1965). "Principal components regression in exploratory statistical research." **Journal American Statistical Association** 60, 234-256.

Obenchain, R. (1980). Comment on "A critique of some ridge regression methods" by G. Smith and F. Campbell. **Journal American Statistical Association** 75, 95-96.

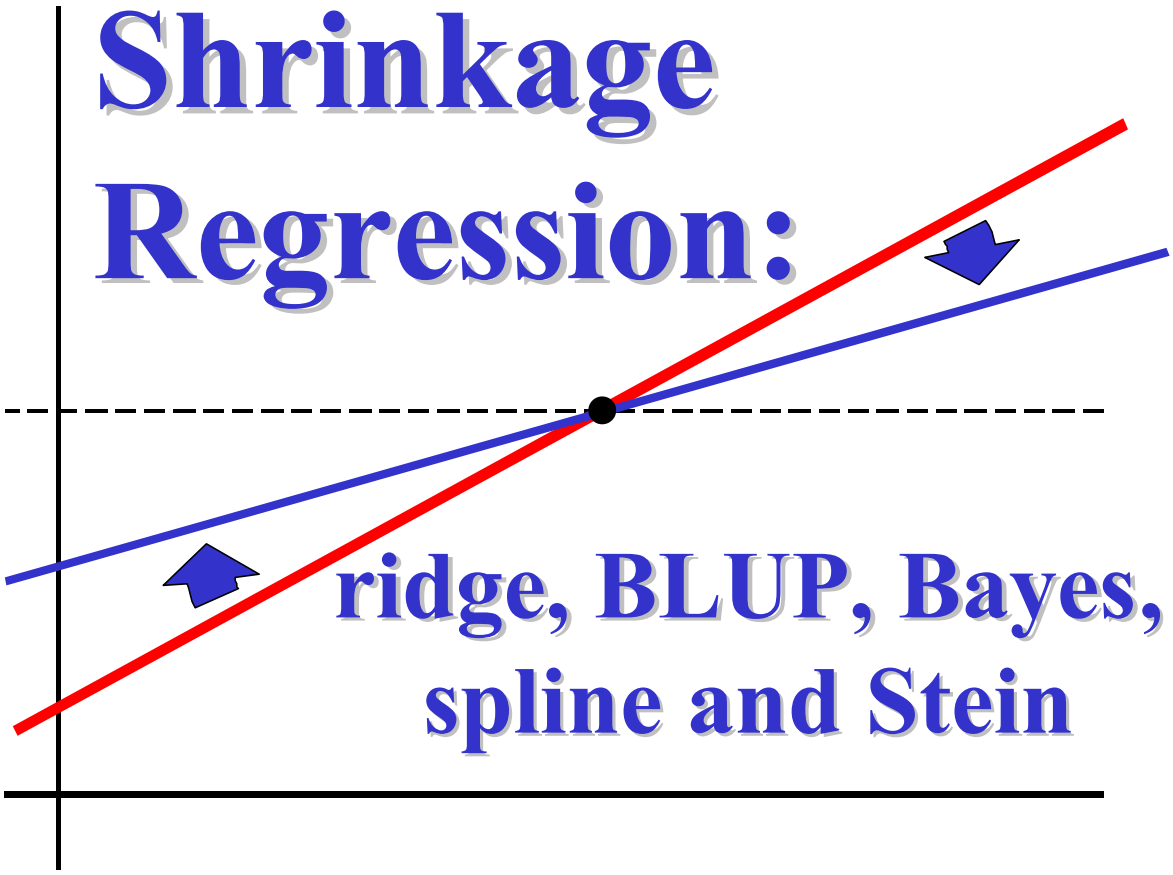
Robinson, G. K. (1991). "That BLUP is a good thing: the estimation of random effects," (with discussion.) **Statistical Science**, 6, 15-51.

Schluster, M. D. (1988). "Unbalanced repeated measures models with structured covariance matrices," **BMDP Statistical Software Manual**, vol.2, 1081-1114. University of California Press, Berkeley.

Searle, S. R. (1979). "Notes on variance component estimation: a detailed account of maximum likelihood and kindred methodology," Biometrics Unit Paper **BU-673-M**, Cornell University (149 pages.)

Searle, S. R. (1988). “Mixed models and unbalanced data: wherefrom, whereat, and whereto?" **Communications in Statistics - Theory and Methods**, 17, 935-968.

Thompson, J. R. (1968). “Some shrinkage techniques for estimating the mean." **Journal American Statistical Association** 63, 113-122.



Chapter 06: Risk Estimation & Simulation

Bob Obenchain, Ph.D.
softR$_X$ freeware
13212 Griffin Run
Carmel, Indiana 46033-8835

Chapter 6: RISK ESTIMATION and SIMULATION

Here in Chapter 6, we explore a variety of normal-theory estimates of the mean-squared-error (MSE) risk resulting from specific shrinkage estimators. We start out in Section §6.1 with what is, perhaps, the single best-known example of an enlighten use for an estimator of shrinkage risk, that of Stein(1973,1981) and Efron and Morris(1976). Section §6.2 displays estimates of relative risk not only in individual components but also in arbitrary linear combinations of **fixed** coefficients, including both bias and range corrections. The developments of Section §6.3 for **random** coefficient models parallel the fixed-coefficient arguments developed in §6.2. Finally, in Section §6.4, we examine Monte-Carlo simulation results that show that risk reduction is easier to actually achieve when coefficients are random than when they are fixed values; after all, the "key" unknown parameters [either a ratio-of-variances or a non-centrality parameter] are very different in these two situations.

6.1 Stein's Unbiased Estimate of Overall Predictive Risk

The early works of Charles Stein [Stein(1955), James and Stein(1961), and Stein(1962)] on normal-theory shrinkage estimation used somewhat tedious mathematical arguments to generate what were, at the time, radically new insights into problems in estimation of three or more mean values under scalar-valued quadratic loss. And the observation of Lindley(1962) that greatly increased contraction (and much lower risk) could result from directing shrinkage toward a linear subspace (of dimension at least 3 less than the original space) certainly helped to start statisticians thinking about the versatility and widespread applicability of shrinkage estimators. Ultimately, the much easier-to-follow arguments of the "unbiased-estimator-of-risk" type described here in Section §6.1 were published by Stein(1973,1981) and Efron and Morris(1976). Our discussion here will closely parallel that Efron and Morris(1976); Jennrich and Oman(1986) also give a highly approachable description of these latter developments, with special attention to their applications in regression.

6.1.1 Contraction Towards a Linear Variety

Let Π represent the orthogonal projection matrix [unique, symmetric, and idempotent; Rao(1973), pp.46-47] for an r-dimensional linear subspace of the P-dimensional space of regression coefficients, β , in our centered multiple regression model of equations { 2.3 } and { 2.4 }. And let β_o represent a $P \times 1$ translation (or shift) vector that lies outside of this r-dimensional subspace [i.e. $\Pi\, \beta_o = 0$]. Elements of a linear variety are then of the general form $\Pi\, \eta + \beta_o$ for some $P \times 1$ vector η.

Two examples of targets for shrinkage are as follows. [i] $\beta_o = 0$ and $\Pi = 1\ 1^T / (1^T 1)$, which is a $P \times P$ matrix with all entries equal to 1/P. This case is a pure projection that defines the one-dimensional linear subspace with all coefficients equal [$\beta_1 = ... = \beta_P$], and was the example Lindley(1962) used in his discussion of Stein(1962). [ii] β_o arbitrary and $\Pi = 0$. Choices of this form yield contraction towards the r=0 dimensional subspace consisting only of the single point, β_o, embedded anywhere within P-dimensional regression coefficient space.

To successfully apply Stein-like contraction methods, we will need to restrict attention to cases where not only the centered regressors matrix is of full rank but also R exceeds r by at least 3. In these cases where R = rank(X) = P, the least squares estimator of β (b^o of equation { 2.6 }) will be uniquely determined.

Now consider, as in Section §2.11, linear hypotheses of the form

$$H: \ (I - \Pi) \beta = \beta_o . \qquad \{ 6.1 \}$$

When this hypothesis holds, β lies entirely within the linear variety (Π , β_o). Technically speaking, equation { 6.1 } actually reads: "The component of β orthogonal to Π equals β_o ." The restricted least squares estimator of β under the hypothesis { 6.1 } is

$$b^H = (X^T X)^{-1} (I - \Pi) W \beta_o + [I - (X^T X)^{-1} (I - \Pi) W (I - \Pi)] b^o , \qquad \{ 6.2 \}$$

where $W = [(I - \Pi) (X^T X)^{-1} (I - \Pi)]^+$, and the corresponding F-ratio test statistic for the hypothesis, H, is

$$F = [(I - \Pi) b^o - \beta_o]^T W [(I - \Pi) b^o - \beta_o] / [(R - r) \cdot s^2] , \qquad \{ 6.3 \}$$

were R = P, R – r = numerator degrees-of-freedom, (N – R – 1) = denominator degrees-of-freedom, and s^2 is the residual-mean-square-for-error of { 2.22 } that is discussed below in section §6.1.2. The non-centrality parameter of this F-statistic is

$$\phi^2(\Pi) = [(I - \Pi) \beta - \beta_o]^T W [(I - \Pi) \beta - \beta_o] / [(R - r) \cdot \sigma^2] . \qquad \{ 6.4 \}$$

6.1.2 Minimum Mean Squared Error Estimation of σ^2

The (N – R – 1) factor in the denominator of $s^2 = y^T (I - H H^T) y / (N - R - 1)$ is widely used (rather than its maximum likelihood value, N, from equation { 5.4 }.) This (N – R – 1) factor makes s^2 an unbiased estimator of σ^2 under normal distribution theory. In fact, s^2 is distributed as the ratio of a central chi-squared random variable divided by its degrees-of-freedom, $\nu = (N - R - 1)$, when the multiple regression model of equations { 2.3 } and { 2.4 } is a correct model and error terms are normally distributed. In this case, the variance of s^2 is $2 \cdot \sigma^4 / \nu$, and the mean-squared-error of $f \cdot s^2$ as an estimator of σ^2, where f is any non-stochastic factor, is

$$\text{MSE}(f \cdot s^2) = \sigma^4 \cdot [f^2 \cdot (\tfrac{2+\nu}{\nu}) - 2 \cdot f + 1] . \qquad \{ 6.5 \}$$

It follows { from equating $\partial \text{MSE}(f \cdot s^2) / \partial f = \sigma^4 \cdot [2 \cdot f \cdot (\frac{2+\nu}{\nu}) - 2]$ to zero and noting that $\partial^2 \text{MSE}(f \cdot s^2) / \partial f^2 = \sigma^4 \cdot 2 \cdot (\frac{2+\nu}{\nu})$ is strictly positive } that the minimum mean-squared-error estimator of σ^2 of the general form $f \cdot s^2$ uses the factor

$$f = (\tfrac{\nu}{\nu+2}) = \tfrac{(N-R-1)}{(N-R+1)} . \qquad \{ 6.6 \}$$

The mean-squared-error of this optimally biased estimator is $2 \cdot \sigma^4 / (\nu + 2)$, which is indeed smaller than the variance, $2 \cdot \sigma^4 / \nu$, of the unbiased estimator, s^2.

In several of the expressions for Stein-like contraction given below, (N – R – 1) factors are counter-balanced by (N – R + 1) factors. These can be interpreted as shrinkage adjustments that provide improved estimation of σ^2 as outlined here in §6.1.2.

6.1.3 Stein Contraction Formulas

Stein-like estimators of β for contraction towards the linear variety (Π, β_o) are of the general form

$$b^S = b^H + \psi(F) \cdot (b^o - b^H) , \qquad \{ 6.7 \}$$

where b^H is the restricted estimator given by equation { 6.2 } and $\psi(F)$ is within a certain class of scalar valued functions of the variance-ratio statistic, F, of equation { 6.3 }. Here, we will consider only the well-known “positive part" form for $\psi(F)$, given by

$$\psi(F) = \max\{ 0 , [1 - \tfrac{K}{F}] \} \quad \text{for } K = \tfrac{(R-r-2) \cdot (N-R-1)}{(R-r) \cdot (N-R+1)} < 1 . \qquad \{ 6.8 \}$$

Now, assuming that one's scalar-valued measure of overall risk in estimation is Predictive Mean Squared Error defined by

$$\text{PMSE}(b) = E[(b - \beta)^T X^T X (b - \beta)] / \sigma^2 , \qquad \{ 6.9 \}$$

Efron and Morris(1976) establish that an unbiased estimator of the PMSE risk associated with the explicitly stochastic shrinkage implied by { 6.8 } is

$$\text{PMSE}(b^S) = \tfrac{(N-R-3) \cdot (R-r)}{(N-R-1)} \cdot F + 2 \cdot r - R \qquad \text{if } F < K , \qquad \{ 6.10 \}$$

$$= R - \tfrac{(R-r-2)^2 \cdot (N-R-1)}{F \cdot (R-r) \cdot (N-R+1)} \qquad \text{otherwise} . \qquad \{ 6.11 \}$$

This PMSE risk estimator is discontinuous at F=K and can be negative, but it yields a truly "enlightening" insight. The numerical values of the unbiased risk estimates of equations { 6.10, 6.11 } can never exceed the PMSE risk, R, of the least squares estimate, b^o, of β. Thus, even though the true PMSE risk of b^s of equation { 6.9 } remains unknown, we do know that the Stein b^s contraction estimator will dominate b^o in terms of PMSE risk.

Other Stein-like results on minimax estimation for scalar valued measures of overall risk (due to Strawderman) are considered in Section §10.x. There we restrict attention to shrinkage to a point ($\Pi = 0$), but we do allow the shape of the shrinkage path to be general (curved), as in the remainder of this chapter.

6.2 Estimates of Shrinkage Risk: Fixed Coefficient Cases

Unfortunately, the elegant arguments of Section §6.1 apply only to the uniform shrinkage case of { 6.7 } when that common shrinkage factor is of the special non-linear and stochastic form given by { 6.8 }. Here in Section §6.2, we discuss estimators of the risk associated with much more general forms of shrinkage of fixed coefficients. But we again impose the (over?) simplifying assumption that all shrinkage factors are to be viewed as non-stochastic. We also consider choice of shrinkage factors to minimize these estimates of risk, as in the minimum C_p approach of Mallows(1973). The risks actually incurred when attempting to optimize risk in the straight-forward (but possibly naive) ways outlined here in Section §6.2 (and in Section §6.3 on random coefficients) are explored using simulation in the last section of this chapter, Section §6.4.

6.2.1 Unbiased Normal-Theory Estimates

We start by displaying estimators for the scaled (relative) risk in individual shrinkage components. We saw in Chapter 4, equation { 4.3 }, that the mean-squared-error risk of $\delta_i \cdot c_i$ as an estimator of the unknown, true (fixed effect) component γ_i is

$$\text{MSE}(\delta_i \cdot c_i) \;=\; \sigma^2 \cdot \delta_i^2/\lambda_i \;+\; (1 - \delta_i)^2 \cdot \gamma_i^2$$

when δ_i is nonstochastic. The corresponding scaled or relative risk is thus $\text{MSE}(\delta_i c_i)/\sigma^2$, a ratio that expresses the risk of a shrunken component estimate as a multiple of the variance of a single observation. The scaled risk can thus be written in the form

$$\tau_{ii} = \text{MSE}(\delta_i c_i) / \sigma^2 \;=\; [\, \delta_i^2 + (1 - \delta_i)^2 \phi_i^2 \,] / \lambda_i \,, \qquad \{ 6.12 \}$$

where the only unknown parameter is the noncentrality (or squared signal-to-noise ratio), $\phi_i^2 = \gamma_i^2 \lambda_i / \sigma^2$. For example, the least-squares solution, $\delta_i = 1$, has completely known relative risk, $\tau_{ii} = 1 / \lambda_i$, because ϕ_i^2 then drops out of equation { 6.12 }. Remember that ϕ_i^2 is the non-centrality parameter of the normal-theory F-ratio for testing the hypothesis that $\gamma_i = 0$. We saw in Chapter 2, equations { 2.16 }, { 2.21 } and { 2.23 }, that this F-ratio is of the form

$$F_i \;=\; \frac{c_i^2\,\lambda_i}{s^2} \;=\; \frac{\nu\cdot r_{yi}^2}{(\,1-R^2\,)}\,,$$

where the denominator degrees-of-freedom are $\nu = N - R - 1$, R is the rank of the centered regressors X matrix, $s^2 = y^T(\,I - H\,H^T\,)\,y\,/\,\nu$ is the least-squares residual-mean-square (unbiased) estimator of σ^2, and $R^2 = r_{y1}^2+r_{y2}^2+...+r_{yR}^2$ is the familiar R-squared statistic. Under normal distribution theory, the R+1 sums-of-squares defined by $(y^Ty)\cdot r_{yi}^2$ for $1 \le i \le R$ and $(y^Ty)\cdot(1-R^2)$ are statistically independent. And the expected value of an F-ratio with 1 numerator degree-of-freedom, $\nu \ge 3$ denominator degrees-of-freedom, and potential for noncentrality only in its numerator is

$$E(\,F_i\,) \;=\; \frac{\nu}{(\,\nu-2)}\cdot[\,\phi_i^2 + 1\,]\,, \qquad \{\,6.13\,\}$$

Johnson and Kotz(1970), equation(3.1), page190. It follows that an estimate of the scaled risk, $MSE(\delta_i c_i)\,/\,\sigma^2$, that is unbiased under normal distribution theory when $\nu \ge 3$ is provided by

$$\widehat{\tau}_{ii} = \{\,2\cdot\delta_i - 1 + (1-\delta_i)^2[\,F_i\cdot(\,\nu-2\,)\,/\,\nu\,]\,\}\,/\,\lambda_i\,. \qquad \{\,6.14\,\}$$

Unbiased estimates can also be developed for the off-diagonal elements of the scaled (relative) mean-squared-error matrix corresponding to equation { 4.2 }. That matrix is

$$T = MSE\,(\,\Delta c\,)\,/\,\sigma^2 \;=\; \Delta^2\,\Lambda^{-1} \;+\; (\,I-\Delta\,)\,\gamma\,\gamma^T\,(\,I-\Delta\,)\,/\,\sigma^2.$$

In fact, arguments parallel to those given above for diagonal elements imply that the corresponding matrix of unbiased estimates, again when $\nu \ge 3$, is of the general form

$$\widehat{T} = (\,\widehat{\tau}_{ij}\,) \;=\; \Lambda^{-1}(\,2\cdot\Delta \;-\; I\,) \;+\; \frac{(\,\nu-2)}{\nu}\cdot(\,I-\Delta\,)\,\Lambda^{-1/2}\,t\,t^T\,\Lambda^{-1/2}(\,I-\Delta\,)\,, \qquad \{\,6.15\,\}$$

where t is the column vector of t-statistics for uncorrelated components with elements defined as in { 2.24 }

$$t = (\,t_{yi}\,) \;=\; \sqrt{\frac{\nu}{(\,1-R^2\,)}}\cdot\;r \qquad \{\,6.16\,\}$$

and $r = (\,r_{yi}\,)$ is again the column vector of principal correlations between the response vector, y , and the columns of the principal axis regressor coordinate matrix, H . Off-diagonal elements of $\widehat{T}$ when $\nu \ge 3$ are thus of the general form:

$$\widehat{\tau}_{ij} = \widehat{\tau}_{ji} = [\;\frac{(1-\delta_i)\cdot r_{yi}}{\lambda_i^{1/2}}\;]\cdot[\;\frac{(1-\delta_j)\cdot r_{yj}}{\lambda_j^{1/2}}\;]\cdot\frac{(\,\nu-2\,)}{(\,1-R^2\,)} \qquad \{\,6.17\,\}$$

for $i \ne j$.

Expression { 6.15 } was first given in Obenchain(1978), equation (3.4); note that this matrix is composed of a known diagonal matrix plus a rank-one matrix defined using the observed t-statistics of the uncorrelated components. The basic building-blocks used to construct this estimator are simply those suggested by maximum-likelihood theory for a multivariate normal

distribution. However, N (the number of observations) from maximum likelihood theory is replaced, here, either by $\nu = (N - R - 1)$ or by $\nu - 2 = (N - R - 3)$. Replacing N with ν is, of course, the well-known adjustment that makes s^2 unbiased for σ^2 . Here, we use $-1 + (\nu - 2) \cdot t_{yi}^2 / \nu$ as our unbiased for ϕ_i^2 and $(\nu - 2) \cdot t_{yi} \cdot t_{yj} / \nu$ as our unbiased for $\phi_i \cdot \phi_j$ when $i \neq j$.

6.2.2 Correct-Range Estimates

As is clear from the relationship $MSE(\Delta c)/\sigma^2 = \Delta^2 \Lambda^{-1} + (I - \Delta)\gamma\gamma^T(I - \Delta)/\sigma^2$, lower bounds on the diagonal elements of the scaled mean-squared-error matrix are

$$\tau_{ii} \;=\; MSE(\delta_i \cdot c_i) / \sigma^2 \;\geq\; \delta_i^2 / \lambda_i \,, \qquad \{ 6.18 \}$$

for $1 \leq i \leq R$. In other words,

the known relative variance is a lower bound for the unknown relative risk

of the shrinkage estimate for each uncorrelated component.

Note that the unbiased estimate of τ_{ii} from equation { 6.14 } [i.e. the element on the diagonal of { 6.15 }] may even be negative when $F_i = t_i^2$ is small and $0 \leq \delta_i < 0.5$. On the other hand, a correct-range estimator of scaled mean-squared-error is given by

$$\tau_{ii}^* \;= \max[\, \widehat{\tau}_{ii} \,,\, \delta_i^2 / \lambda_i \,] \,, \qquad \{ 6.19 \}$$

$$= \delta_i^2 / \lambda_i + \max[\, 0 \,,\, \tfrac{(\nu-2)}{\nu} \cdot F_i \;- 1 \,] \cdot (1 - \delta_i)^2 / \lambda_i$$

where $\widehat{\tau}_{ii}$ is the unbiased estimator of { 6.14 } and { 6.15 } .

When either $\nu = N - R - 1 \leq 2$ or the i-th principal regressor correlation with the response, r_{yi} , is sufficiently close to zero that $(\nu - 2) \cdot F_i / \nu$ is less than 1, the estimated scaled mean-squared-error of $\delta_i \cdot c_i$ will continually decrease as δ_i decreases, reaching a minimum of 0 at $\delta_i = 0$.

Otherwise, when r_{yi} is large enough to make $(\nu - 2) \cdot F_i / \nu$ greater than 1, the estimated scaled mean-squared-error of $\delta_i \cdot c_i$ will reach a strictly positive minimum value at a strictly positive value of δ_i. Specifically, in this case, $(\nu - 2) \cdot F_i / \nu = 1 + f$ for some strictly positive factor, $f > 0$. Then, taking derivatives as in equations { 4.4 } and { 4.5 }, minimum estimated risk of $f / [\, (1+f) \cdot \lambda_i \,]$ is achieved at the strictly positive value, $\delta_i = f / (1+f) = 1 - 1/(1+f)$.

It turns out that both the cases where $|r_{yi}|$ is small and those where $|r_{yi}|$ is large can be summarized quite simply, as detailed next.

6.2.3 Shrinkage Factors Minimizing Scaled Risk Estimates

The shrinkage factor value, δ_i^* , that minimizes both the unbiased and the correct-range estimates, $\widehat{\tau}_{ii}$ of { 6.14 } and τ_{ii}^* of { 6.19 } , of the scaled mean-squared-error in $\delta_i \cdot c_i$ is

$$\begin{aligned} \delta_i^* &= 1 - \frac{\nu}{(\nu-2)\cdot F_i} && \text{if } \frac{(\nu-2)\cdot F_i}{\nu} > 1 , \\ &= 0 && \text{otherwise,} \end{aligned} \qquad \{ 6.20 \}$$

where $\nu = N - R - 1$. Thus, by its very definition, $\delta_i^* \equiv 0$ when $r_{yi}^2 = 0$. Otherwise, δ_i^* approaches 1 as R^2 approaches 1 because $F_i = c_i^2 \cdot \lambda_i / s^2$ becomes arbitrarily large in this limiting case.

Note that δ_i^* of { 6.20 } would also result from imposing a non-negativity restriction, max(0, $\widehat{\phi}^2$), on an otherwise unbiased estimator of ϕ_i^2 and plugging that value into $\delta_i^{MSE} = \phi_i^2 / (1 + \phi_i^2)$ of { 4.6 } . By way of contrast, the normal-theory maximum likelihood estimate of $\phi_i^2 = \gamma_i^2 \cdot \lambda_i / \sigma^2$ is directly proportional to $F_i = c_i^2 \cdot \lambda_i / s^2$, as established in section §5.2.1. In fact, except for using $\nu = (N - R - 1)$ instead of the maximum-likelihood value of N in the denominator of $s^2 = y^T (I - H H^T) y / \nu$, the normal-theory maximum-likelihood estimator of δ_i^{MSE} is of the form

$$\widehat{\delta}_i^{MSE} = F_i / (1 + F_i) , \qquad \{ 6.21 \}$$

without regard to whether the numerical size of F_i is ≤ 1 or ≥ 1. Note that the resulting product, $\widehat{\delta}_i^{MSE} \cdot c_i$, corresponds to the Thompson(1968) "cubic" estimator of the true component, γ_i . The shrinkage estimate of { 6.21 } will be called asymptotic maximum likelihood because N is replaced by $\nu = N - R - 1$.

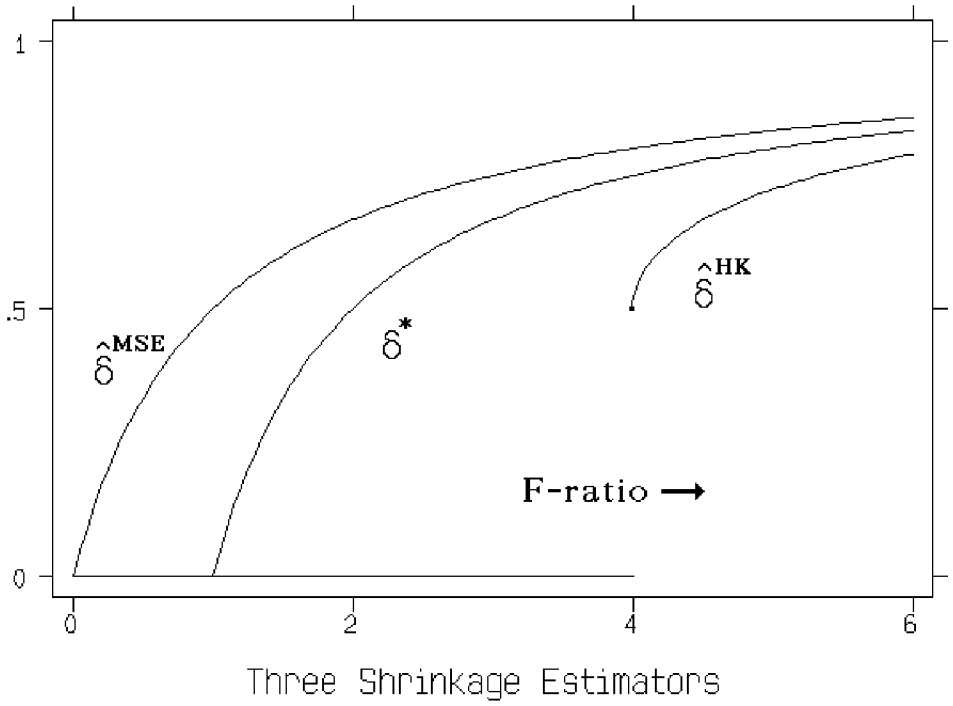
For comparison with { 6.20 } and { 6.21 } , we remark that Hemmerle(1975) showed that the heuristic "fixed-point" iteration of Hoerl and Kennard(1970a,b) converges to the almost drastic shrinkage value:

$$\begin{aligned} \widehat{\delta}_i^{HK} &= 0 && \text{if } 0 \leq F_i \leq 4, \\ &= (1 + \sqrt{1 - 4 \cdot F_i^{-1}}) / 2 && \text{otherwise.} \end{aligned} \qquad \{ 6.22 \}$$

Figure 6.1 below illustrates the relative extents of shrinkage implied by equations { 6.20 }, { 6.21 } and { 6.22 } when ν is very large. In this limiting case, δ_i^* of { 6.20 } is approximately max[0 , $1 - F_i^{-1}$] , as in the minimum C_p approach of Mallows(1973). Note, in particular, that the minimum estimated risk shrinkage estimator of { 6.20 } yields considerably

more shrinkage than the maximum-likelihood solution of { 6.21 } when F_i is small. But neither of these shrinkage solutions is nearly as drastic as the Hoerl-Kennard-Hemmerle solution over the $1 < F_i < 4$ range.

Figure 6.1 Three Shrinkage Extent Estimators



As we remarked when we first wrote equation { 6.12 }, the noncentrality, ϕ_i^2, is the key unknown ingredient defining the scaled mean-squared-error risk, τ_{ii} , corresponding to different numerical values for the non-stochastic shrinkage factor, δ_i . And we wrote equations { 6.14 }, { 6.19 } and { 6.20 } in forms that also emphasize the importance of one's estimate of this noncentrality. The three primary estimates of ϕ_i^2 we have considered here in Section §6.2 are...

Asymptotic Maximum Likelihood: $\quad \phi_i^2$ estimate $= F_i$

Normal-Theory Unbiased [$\nu \geq 3$]: $\quad \phi_i^2$ estimate $= \frac{(\nu-2)}{\nu} \cdot F_i - 1$

Correct-Range Modification: $\quad \phi_i^2$ estimate $= \max[\, 0 \,,\, \frac{(\nu-2)}{\nu} \cdot F_i - 1 \,]$

As we shall see below in Section §6.4, the mean-squared-error risk of the unbiased estimate of ϕ_i^2 uniformly dominates that of the normal-theory maximum-likelihood estimate; in fact, its risk is smaller by almost a factor of 10 when $\nu = 3$. In turn, the mean-squared-error risk of the correct range estimate of ϕ_i^2 uniformly dominates that of the unbiased estimate; but differences in risk are quite small here unless the true noncentrality is small.

Unfortunately, as we shall also see in Section §6.4, a very good estimator of ϕ_i^2 does not necessarily yield a good estimator of $\delta_i^{MSE} = \phi_i^2 / (1 + \phi_i^2)$ of { 4.6 } , let alone assure that the product of c_i times that [estimated δ_i^{MSE}] will be a good shrinkage estimator of the true γ_i .

6.2.4 The Estimated Risk in Arbitrary Linear Combinations

We will write MSE($\alpha^T b^\bigstar$) / σ to denote the scaled (or relative) mean-squared-error of $\alpha^T b^\bigstar = \alpha^T G \Delta c$ as an estimator of $\alpha^T \beta$, where the α vector defines an arbitrary linear combination of generalized shrinkage regression estimates, $b^\bigstar$, and G is the direction cosines matrix of { 2.8 }. Geometrically speaking, this is simply the relative MSE parallel to $\pm \alpha$ in P-dimensional regression coefficient space. Algebraically, $\alpha^T b^\bigstar = \alpha^T G \Delta c$ is simply a known linear combination of the shrunken components, Δc . As a result, the scaled MSE of $\alpha^T b^\bigstar$ is unbiasedly estimated by forming the inner product, $\alpha^T G \widehat{T} G^T \alpha$, where $\widehat{T}$ of { 6.16 } is the scaled MSE of Δc as an estimator of the uncorrelated components vector, γ .

One way to generate correct-range estimates of MSE($\alpha^T b^\bigstar$) / σ would then be to replace the diagonal elements of $\widehat{T}$ in $\alpha^T G \widehat{T} G^T \alpha$ with the τ_{ii}^* of equation { 6.20 }. On the other hand, numerically smaller estimates with correct-range can sometimes result from retaining the $\widehat{\tau}_{ii}$ diagonal elements of equation { 6.16 } but taking one's estimate of MSE($\alpha^T b^\bigstar$) / σ to be of the form:

$$\max(\alpha^T G \widehat{T} G^T \alpha , \alpha^T G \Delta^2 \Lambda^{-1} G^T \alpha) .$$

Unbiased estimates of the entire scaled mean-squared-error matrix, MSE($b^\bigstar$) / σ , of generalized shrinkage regression estimates are of the form $G \widehat{T} G^T$ for the $\widehat{T}$ of { 6.16 }. And replacing the diagonal elements of $\widehat{T}$ with the τ_{ii}^* of equation { 6.20 } yields a natural choice for a correct-range estimate of this relative risk matrix. The diagonal elements of this $G \widehat{T}^* G^T$ matrix are plotted in a TRACE display by my RXridge software. Similarly, the scaled (or relative) version of the excess-mean-squared-error-matrix for least-squares minus ridge, EMSE of { 4.25 }, is estimated by $G (\Lambda^{-1} - \widehat{T}^*) G^T$. The eigenvalues of this estimated relative EMSE risk matrix are also plotted in a TRACE display by RXridge, along with a TRACE of the inferior-direction associated with any negative eigenvalue of the relative excess-mean-squared-error matrix.

6.2.5 Mallows-like Estimates of Predictive Mean-Squared-Error

Mallows(1973) defined the Predictive MSE Risk of a shrinkage estimator, $b^\bigstar$, of β to be

$$PMSE(b^\bigstar) = 1 + \frac{1}{\sigma^2} \cdot E[(X b^\bigstar - X \beta)^T (X b^\bigstar - X \beta)] , \qquad \{ 6.23 \}$$

$$= 1 + \frac{1}{\sigma^2} \cdot \sum_{i=1}^{R} \lambda_i \cdot \mathrm{MSE}(\delta_i \cdot c_i) ,$$

$$= 1 + \frac{1}{\sigma^2} \cdot \sum_{i=1}^{R} \lambda_i \cdot \tau_{ii} ,$$

where the τ_{ii} of { 6.18 } are the diagonal elements of the MSE risk matrix for $b^{\bigstar}$. components. And Mallows' estimator of this risk is of the form

$$C(b^{\bigstar}) = (N - R - 1) \cdot \frac{RMS^{\bigstar}}{RMS^{o}} - N + 2 \cdot \sum_{i=1}^{R} \delta_i + 2 , \qquad \{ 6.24 \}$$

where $RMS^{\bigstar}$ and RMS^{o} are the residual-mean-squares corresponding to the shrinkage estimate, $b^{\bigstar}$, and the least-squares estimate, b^{o} , of β from equation { 3.5 }. We can rewrite { 6.24 } using the relationship

$$\frac{RMS^{\bigstar}}{RMS^{o}} = 1 + \frac{r^T (I-\Delta)^2 r}{(1-R^2)} = 1 + \frac{t^T (I-\Delta)^2 t}{(N-R-1)} , \qquad \{ 6.25 \}$$

as

$$C(b^{\bigstar}) = t^T (I - \Delta)^2 t - R + 2 \cdot \sum_{i=1}^{R} \delta_i + 1 , \qquad \{ 6.26 \}$$

$$= 1 + \frac{(N-R-1)}{(N-R-3)} \cdot \mathrm{trace}[\Lambda^{1/2} \hat{T} \Lambda^{1/2}] - 2 \cdot \frac{(2 \cdot \sum \delta_i - R)}{(N-R-3)},$$

where the $\hat{T}$ matrix contains the unbiased risk estimates of { 6.16 }. Thus Mallows' estimator of Predictive MSE is biased; an unbiased estimator is provided by

$$C^{U}(b^{\bigstar}) = 1 + \mathrm{trace}[\Lambda^{1/2} \hat{T} \Lambda^{1/2}]$$

$$= (N - R - 3) \cdot \frac{(RMS^{\bigstar} - RMS^{o})}{RMS^{o}} - R + 2 \cdot \sum_{i=1}^{R} \delta_i + 1 . \qquad \{ 6.27 \}$$

Note that this unbiased estimate of Predictive MSE is minimized when the shrinkage factors coincide with the δ_i^* choices of { 6.21 }.

Usage of Mallows' C-statistic estimates of Predictive MSE has a strong tradition in the area of regressor variable subset selection. Thus we will return to the general topic of risk estimation in our discussion of computationally intensive methods, Chapter §10. Mallows(1973) suggested a way of superimposing Predictive MSE estimates for shrinkage regression, { 6.25 }, on the same "C_p versus p" graph that would be used for regressor variable subset selection, where p denotes the rank of a subset that includes the constant term ($1 \leq p \leq R + 1$.) Mallows' proposal is equivalent to plotting $C(b^{\bigstar})$ versus $\sum \delta_i^2 + 1$. Arguments too involved to detail

here suggest that it would be much more appropriate to plot C($b^{\star}$) versus $\sum \delta_i + 1$ if one's objective were to superimpose the resulting curve on top of the "C_p versus p" plot for subset selection.

6.3 Estimates of Shrinkage Risk: Random Coefficient Cases

Here in Section §6.3, we discuss estimators of the risk associated with general forms of non-stochastic shrinkage of random coefficients. We will see that there are so many parallels here with the fixed-effect results of Section §6.2 that we can omit most details.

Suppose we start with an unbiased estimate, c , of a random effect, γ , that is subject to additive noise with variance σ^2 . In other words, the conditional expected value of c given γ would then be $E(c \mid \gamma) = \gamma$ and the conditional variance of c given γ would be $V(c \mid \gamma) = \sigma^2$. Then, exactly as in the fixed-effect derivation of equation { 4.3 }, the conditional mean-squared-error risk of $\delta \cdot c$ as an estimator of the given γ would be

$$MSE(\delta \cdot c \mid \gamma) = E[(\delta \cdot c - \gamma)^2 \mid \gamma] = \sigma^2 \cdot \delta^2 + (1 - \delta)^2 \cdot \gamma^2$$

when δ is nonstochastic. If the expected value of γ is zero and the variance of γ is σ_γ^2 , the resulting unconditional mean-squared-error risk of $\delta \cdot c$ as an estimator of the unknown, random γ would then be

$$MSE(\delta \cdot c) = E[MSE(\delta \cdot c \mid \gamma)] = \sigma^2 \cdot \delta^2 + (1 - \delta)^2 \cdot \sigma_\gamma^2 , \qquad \{ 6.28 \}$$

again assuming that δ is a known constant. The corresponding scaled (or relative) risk is thus

$$\tau = MSE(\delta \cdot c) / \sigma^2 = \delta^2 + (1 - \delta)^2 \phi^2 \qquad \{ 6.29 \}$$

where the only unknown parameter is the variance ratio $\phi^2 = \sigma_\gamma^2 / \sigma^2$.

Then, if one had an unbiased estimate, s^2 , of σ^2 based upon ν degrees-of-freedom that was independent of c, the following variance – ratio F-statistic would be the Normal-theory maximum likelihood estimator of ϕ^2 :

$$F = \frac{c^2}{s^2} . \qquad \{ 6.30 \}$$

In fact, just as in { 6.13 }, the Normal-theory expected value of this F-ratio when $\nu \geq 3$ would be

$$E(F) = \frac{\nu}{(\nu - 2)} \cdot [\phi^2 + 1] . \qquad \{ 6.31 \}$$

Because the above formulas are of the exact same functional form as the corresponding fixed-effect results of Section §6.2, we will not need to repeat details here on MSE risk estimation and on the shrinkage factor values that minimize those estimates of risk. All you need to remember is that (i) the random-effect signal standard deviation, σ_γ , plays the same role as the

fixed-effect expected signal, γ_i , and that (ii) the random-effect noise standard deviation is denoted by σ (or s) rather than by σ / $\lambda_i^{1/2}$ (or s / $\lambda_i^{1/2}$.)

Our three primary estimates of ϕ^2 are, again...

Maximum Likelihood: ϕ^2 estimate = F

Normal-Theory Unbiased [$\nu \geq 3$]: ϕ^2 estimate = $\frac{(\nu-2)}{\nu} \cdot F - 1$

Correct-Range Modification: ϕ^2 estimate = max[0 , $\frac{(\nu-2)}{\nu} \cdot F - 1$]

6.4 Monte-Carlo Risk Simulation

Three very different sorts of approaches to development of mean-squared-error risk profiles for shrinkage regression estimators have been discussed in statistical literature and applied to a wide variety of different "realizable" shrinkage estimators. These three approaches are:

(i) Derivation of exact, analytical expressions. For example, see Dwivedi, Srivastava, and Hall(1980) and Hemmerle and Carey(1981).

(ii) Numerical integration and approximation. For example, this approach was used by Thompson(1968), Lawless(1975) and Kadiyala(1980).

(iii) Monte-Carlo simulation techniques. This approach has been used by a large number of authors; for example, see Newhouse and Oman(1971), McDonald and Galarneau(1975), Hoerl, Kennard and Baldwin(1975), Lawless and Wang(1976), Obenchain(1975b,1976), Dempster, Schatzoff, and Wermuth(1977), Gunst and Mason(1977), Yancey and Judge(1977), Hemmerle and Brantley(1978), Wichern and Churchill(1978), Goto(1979), Goto and Matsubara(1979), Gibbons(1981), Hoerl, Schuenemeyer and Hoerl(1986), Jennrich and Oman(1986), Krishnamurthi and Rangaswamy(1987,1990).

With today's widespread availability of software simulation tools and high-speed computing hardware (see Chapter 15), Monte-Carlo techniques can be particularly attractive and versatile. For example, simulation can be used to show how seemingly "minor" changes in formulas for the extent of shrinkage can result in "major" changes in implied risk profiles. And, of course, the simulation approach is quite easily adapted for study of non-normal error distributions. Draper and Van Nostrand(1977) suggest that many published simulation studies may be "biased" [intentionally or unintentionally, possibly in quite subtle ways] in favor of shrinkage methods. Be that as it may, the fact remains that the vast majority of published simulation studies point quite strongly toward a rather "optimistic" point-of-view:

Several different shrinkage regression methods all compare quite favorably with least-squares in terms of mean-squared-error risk.

6.4.1 Simulated Risk for Fixed Coefficient Models

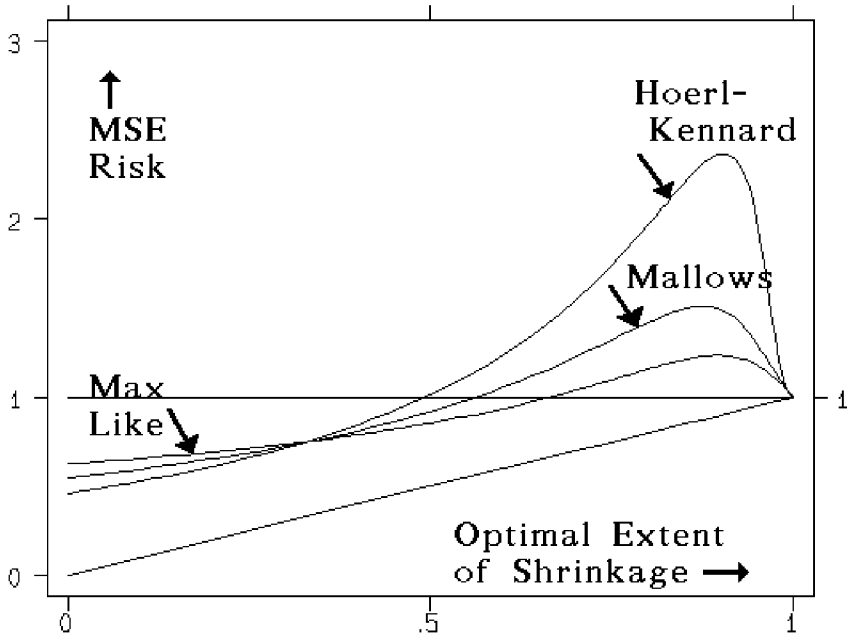
Figures 6.2, 6.3 and 6.4 display simulated mean-squared-error risk profiles for the three fixed coefficient shrinkage estimators whose relative extents of shrinkage were displayed in Figure 6.1. These estimators are: maximum likelihood shrinkage as in { 6.22 }, Mallows' minimum estimated-risk from { 6.21 }, and Hoerl-Kennard-Hemmerle drastic-shrinkage from { 6.23 }. In each of these figures, the horizontal axis corresponds to a range of optimal extents for shrinkage, $\delta_i^{MSE} = \phi_i^2 / (\phi_j^2 + 1)$ from equation { 4.6 }. [For example, $\delta_i^{MSE} = 0$ when $\phi_j^2 = 0$ because the i-th true component is $\gamma_i = 0$. And $\delta_i^{MSE} = 1$ is the limiting case where the ϕ_i^2 noncentrality of $F_i = c_i^2 \lambda_i / s^2$ of { 2.22 } approaches infinity.] The profile of Figure 6.2 results when only one degree-of-freedom is available for estimating the error variance, σ^2. Figure 6.3 depicts the case where error d.f.=5. And Figure 6.4 describes the limiting case where σ^2 is known (error d.f.=∞.)

The results displayed below in Figures 6.2, 6.3 and 6.4 (and listed in the tabulations below each figure) were generated, as described in Chapter §15, using 5 million Monte-Carlo replications with my RXmsesim.EXE software for IBM-compatible personal computers. Results for different optimal-shrinkage extents are as smoothly self-consistent (highly positively correlated) as is possible in the sense that they all were generated using the exact same sequence of pseudo-random, normally-distributed variates. And comparison of the theoretical optimal-shrinkage values for the first column of each tabulation with the corresponding last column (the simulated risk resulting from shrinkage of exactly optimal extent as in { 4.7 }) should convince you that the risk estimates in these tabulations are accurate to 3 decimal places.

Note, in particular, that most risk differences between the known error-variance case ($\nu = \infty$ of Figure 6.4) and unknown error-variance cases (of Figures 6.2 and 6.3) are really rather small numerically. We can summarize our Monte-Carlo simulation findings for fixed coefficient cases as follows:

> The almost-drastic-shrinkage suggestion of Hoerl and Kennard(1970a,b) and Hemmerle(1975) performs quite well when drastic shrinkage is appropriate (δ^{MSE} is nearly zero), reducing mean-squared-error risk by as much as 86%. But this same tactic can also increase risk by as much as 143% when drastic shrinkage is inappropriate (δ^{MSE} in the 0.85 to 0.90 range.)
>
> The minimum-estimated-risk suggestion [like that of Mallows(1973)] also performs well when drastic shrinkage is appropriate, reducing mean-squared-error risk by as much as 68%. But this tactic can also increase risk by as much as 49% when δ^{MSE} is approximately 0.85 to 0.875.
>
> **Continued...**

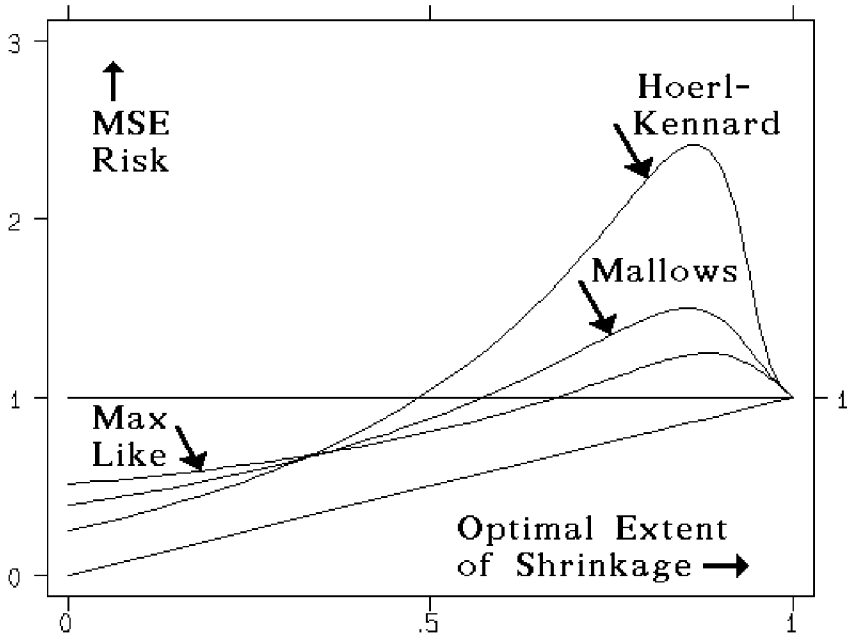
Figure 6.2 Simulated Fixed Coefficient Risk when Error Degrees-of-Freedom = 1.



Simulated Fixed Coefficient Risks for Error Degrees-of-Freedom = 1 :

delta	H-K-H	Mallows	maxLike	minMSE
0.0000	0.4577	0.5462	0.6248	0.0000
0.1000	0.5281	0.5955	0.6553	0.1000
0.2000	0.6128	0.6541	0.6915	0.2001
0.3000	0.7168	0.7249	0.7350	0.3001
0.4000	0.8474	0.8115	0.7881	0.4001
0.5000	1.0158	0.9195	0.8538	0.5001
0.6000	1.2399	1.0557	0.9363	0.6001
0.7000	1.5472	1.2265	1.0394	0.7001
0.8000	1.9682	1.4210	1.1588	0.8001
0.9000	2.3654	1.4947	1.2348	0.9001
0.9900	1.0597	1.0520	1.0470	0.9900

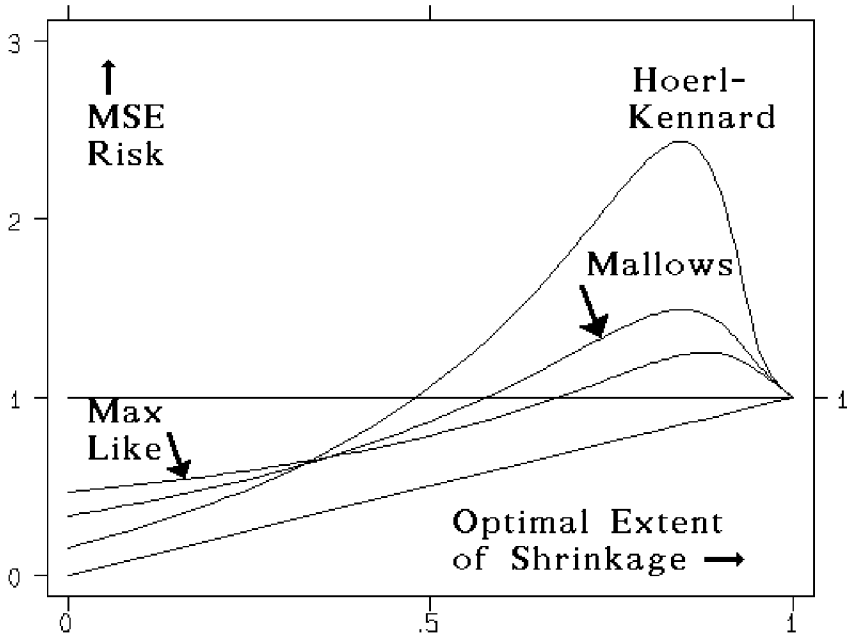
Figure 6.3 Simulated Fixed Coefficient Risk when Error Degrees-of-Freedom = 5.



Simulated Fixed Coefficient Risks for Error Degrees-of-Freedom = 5 :

delta	H-K-H	Mallows	maxLike	minMSE
0.0000	0.2456	0.3932	0.5094	0.0000
0.1000	0.3462	0.4589	0.5489	0.1000
0.2000	0.4673	0.5366	0.5957	0.2001
0.3000	0.6154	0.6295	0.6517	0.3001
0.4000	0.8008	0.7419	0.7198	0.4001
0.5000	1.0374	0.8795	0.8037	0.5001
0.6000	1.3460	1.0483	0.9079	0.6000
0.7000	1.7502	1.2492	1.0357	0.7000
0.8000	2.2306	1.4485	1.1769	0.8000
0.9000	2.2908	1.4387	1.2428	0.9000
0.9900	1.0370	1.0354	1.0339	0.9900

Figure 6.4 Simulated Fixed Coefficient Risk when the Variance is Known.



Simulated Fixed Coefficient Risks for Known Variance (Degrees-of-Freedom = ∞) :

delta	H-K-H	Mallows	maxLike	minMSE
0.0000	0.1531	0.3335	0.4671	0.0000
0.1000	0.2676	0.4057	0.5099	0.1000
0.2000	0.4056	0.4908	0.5604	0.2000
0.3000	0.5749	0.5923	0.6208	0.3000
0.4000	0.7865	0.7145	0.6942	0.4000
0.5000	1.0565	0.8632	0.7845	0.4999
0.6000	1.4058	1.0435	0.8963	0.5999
0.7000	1.8515	1.2540	1.0329	0.6999
0.8000	2.3293	1.4514	1.1820	0.7999
0.9000	2.1455	1.4141	1.2446	0.8999
0.9900	1.0322	1.0312	1.0303	0.9899

The normal-theory, maximum-likelihood approach of Obenchain(1975,1981,1984) shrinks least aggressively even when drastic shrinkage is appropriate, reducing mean-squared-error risk by only 47% to 53%. But this tactic also never increases risk by more than 23% to 25% ...even in the least favorable situation where δ^{MSE} is approximately 0.875 to 0.90.

Both the minimum-estimated-risk and the maximum likelihood approaches have the desirable property that they can result in a larger percentage-wise decrease in risk than their own worst-case increase in risk. And maximum likelihood limits its worst-case increase in risk to only about 25% above the minimax least-squares level.

The table below, Table 6.1, suggests that the unbiased and correct-range estimates of fixed-coefficient noncentrality, $\phi^2 = \gamma^2 \lambda / \sigma^2$, have uniformly smaller mean-squared-error risk than does the asymptotic maximum likelihood (F-ratio) estimate. These risks are well defined only when the degrees-of-freedom for error are at least 5; the variance of the F-ratio used in all three noncentrality estimates is

$$V(F_i) = \frac{2\cdot\nu^2\cdot[\,\phi^4+(1+2\cdot\phi^2)\cdot(\nu-1)\,]}{(\nu-2)^2\cdot(\nu-4)}, \qquad \{ 6.32 \}$$

Johnson and Kotz(1970), equation (3.3), page 190.

Table 6.1 Simulated MSE Risk in Noncentrality Estimation when Coefficients are Fixed.

First Row Label: T => Theoretical risk of maximum likelihood
M => simulated risk of Maximum likelihood
U => simulated risk of Unbiased estimate
C => simulated risk of Correct range estimate

Second Row Label: Degrees-of-Freedom for Error (noise) estimation
First Column Label: MSE Optimal Shrinkage Factor, $\phi^2 / (\phi^2 + 1)$
Second Column Label: Squared Signal/Noise Ratio, $\phi^2 = \gamma^2 \lambda / \sigma^2$

		0.0000	0.2000	0.4000	0.6000	0.8000	0.9900
		0.0000	0.5000	0.8167	1.2250	2.0000	9.9500
T	5	25.00	37.04	58.78	108.50	307.67	63451
M	5	24.75	36.56	57.61	105.55	296.07	58467
U	5	7.912	11.95	19.14	35.44	99.84	19401
C	5	7.451	11.28	18.20	34.22	98.68	19401
T	14	4.900	6.785	10.011	16.775	39.567	3684.4

M	14	4.913	6.805	10.039	16.820	39.671	3693.6
U	14	2.608	3.924	6.173	10.879	26.667	2483.2
C	14	2.193	3.323	5.344	9.825	25.770	2483.2
T	29	3.737	5.076	7.332	11.949	26.612	1489.5
M	29	3.734	5.064	7.315	11.922	26.560	1487.5
U	29	2.237	3.356	5.248	9.118	21.398	1228.5
C	29	1.832	2.769	4.442	8.100	20.550	1228.5
T	99	3.191	4.277	6.094	9.752	20.910	651.98
M	99	3.184	4.268	6.082	9.735	20.880	651.26
U	99	2.059	3.089	4.814	8.287	18.881	616.27
C	99	1.659	2.511	4.021	7.290	18.063	616.27
T	∞	3.000	4.000	5.667	9.000	19.000	399.000
M	∞	3.002	4.005	5.673	9.008	19.010	398.936
U	∞	2.003	3.005	4.673	8.007	18.008	397.924
C	∞	1.605	2.430	3.885	7.018	17.202	397.924
–	—						
		0.0000	0.2000	0.4000	0.6000	0.8000	0.9900
		0.0000	0.5000	0.8167	1.2250	2.0000	9.9500

By comparing the true and simulated MSE risks of the F-ratio estimate (the rows marked T and M in the above table), you observe that the simulation results of Table 6.1 are apparently accurate to 2 or 3 decimal places, at least when the degrees-of-freedom for error are ≥ 14. The simulation results for degrees-of-freedom $= 5$ are somewhat less accurate even though they too are based upon 5 million Monte-Carlo replications. On the other hand, because our simulation strategy was again to use the exact same sequence of pseudo-random, normal deviates in evaluating all three ϕ_i^2 estimates, these simulation results are as smoothly self-consistent (highly positively correlated) as is possible. Thus the simulation results of Table 6.1 strongly support my conjecture that the unbiased and correct-range estimates of fixed-coefficient noncentrality, $\phi^2 = \gamma^2 \lambda / \sigma^2$, have uniformly smaller mean-squared-error risk than does the asymptotic maximum likelihood (F-ratio) estimate under Normal distribution-theory.

The "bad news" here about noncentrality estimation is that the MSE risk superiority of the correct-range estimate of ϕ_i^2 does not necessarily translate into a superior shrinkage estimate of γ_i. Specifically, added shrinkage results from using the correct-range estimate of ϕ_i^2 in $\widehat{\delta}_i = \widehat{\phi}_i^2 / (\widehat{\phi}_i^2 + 1)$ than when using the asymptotic maximum likelihood (F-ratio) estimate of ϕ_i^2. This additional shrinkage yields an estimator of γ_i whose risk profile (i) is more extreme than the Hoerl-Kennard option when the degrees-of-freedom for error are 3 or 4, (ii) lies somewhere "between" the Mallows and the Hoerl-Kennard options when the degrees-of-freedom for error exceed 5, and (iii) becomes virtually indistinguishable from the Mallows profile when the degrees-of-freedom for error exceed 99 (just as in our arguments on minimization of τ_{ii}^* of { 6.19 }.)

In summary, then, the correct-range modification to the unbiased estimate of ϕ_i^2 apparently does yield an improved estimate of ϕ_i^2 . Furthermore, because ϕ_i^2 is the only unknown in expression { 6.12 } for the relative risk of nonstochastic shrinkage (and because ϕ_i^2 is in the numerator of that expression), it follows that the correct-range estimate of ϕ_i^2 also yields superior estimates of the relative risk associated with nonstochastic shrinkage. However, remember that the shrinkage factor values that actually minimize these improved risk estimates are stochastic.

Simulation results such as those of Figures 6.2, 6.3 and 6.4 for the Hoerl-Kennard, Mallows, and asymptotic maximum likelihood approaches specifically account for the stochastic nature of the shrinkage being applied. And simulation results account not only for (i) the stochastic nature of the shrinkage factor estimate ($\widehat{\delta}_i^{HK}$, δ_i^* or $\widehat{\delta}_i^{MSE}$) but also for (ii) the correlation between this factor estimate and the corresponding component, c_i , of the least-squares, fixed-effect regression coefficient vector. At least in my opinion, these simulation results strongly favor the (relatively conservative) asymptotic maximum likelihood approach.

6.4.2 Simulated Risk for Random Coefficient Models

Figures 6.5, 6.6 and 6.7 display simulated mean-squared-error risk profiles in random coefficient models for the same three shrinkage estimators described in Figures 6.2, 6.3 and 6.4 for fixed-coefficient models. Again, these estimators are: asymptotic maximum likelihood shrinkage as in { 6.22 }, Mallows' minimum estimated-risk as in { 6.21 }, and Hoerl-Kennard-Hemmerle almost-drastic-shrinkage as in { 6.23 }. The horizontal axis again corresponds to a range of optimal extents for shrinkage, $\delta^{MSE} = \phi^2 / (\phi^2 + 1)$ from $\delta^{MSE} = 0$ to $\delta^{MSE} = 0.99$, where the key parameter is now the variance-ratio, $\phi^2 = \sigma_\gamma^2 / \sigma^2$, associated with a single random-coefficient.

The risk profile of Figure 6.5 results when only one degree-of-freedom is available for estimating the error variance, σ^2. Figure 6.6 depicts the case where the error d.f.=5. And Figure 6.7 describes the limiting case where σ^2 is known (error d.f.=∞.) These simulation results were also generated using at least 5 million Monte-Carlo replications with my RXmsesim.EXE software for IBM-compatible personal computers (see Chapter §15) and, again, appear to be accurate to three decimal places.

Risk results for $\delta^{MSE} = 0$ should be identical to those for fixed-coefficient models; after all, fixed-effect and random-effect models are equivalent in this limiting, special case. Thus it is somewhat reassuring that all pairs of fixed-effect and random-effect MSE risk estimates for all $\delta^{MSE} = 0$ cases differ by no more than 0.0009.

Perhaps the most striking observation that results from comparing the random-coefficient risk profiles (Figures 6.5, 6.6 and 6.7) with the corresponding fixed-coefficient risk profiles (Figures 6.2, 6.3 and 6.4) is that...

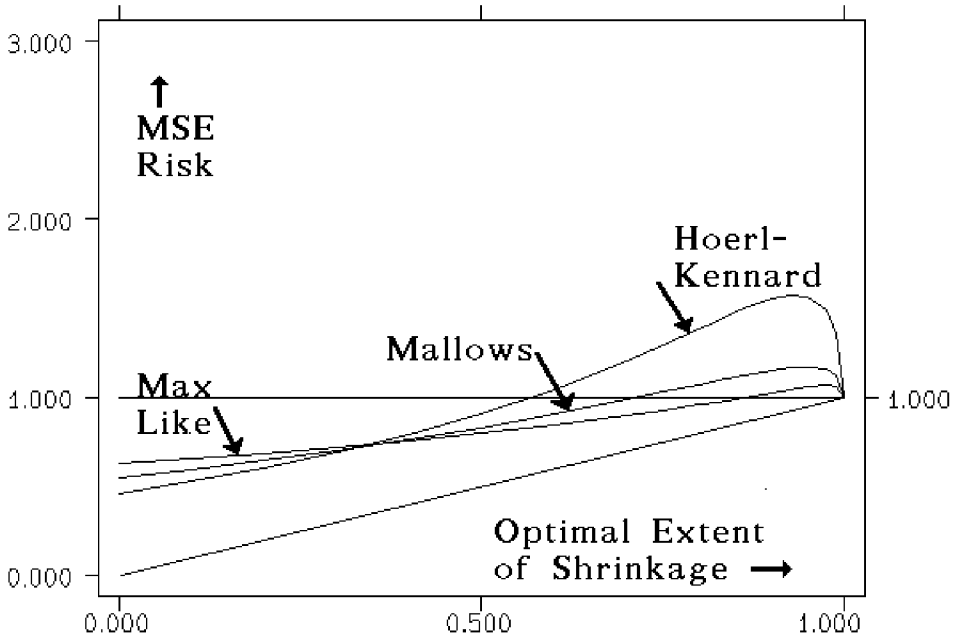
Shrinkage estimation does a much better job of either reducing and/or limiting increases in MSE risk when coefficients vary randomly than when they are fixed.

Specifically, the least-favorable extend of optimal shrinkage tends to fall roughly in the 0.80 to 0.90 range when coefficients are fixed. When coefficients are random, the least-favorable extents for shrinkage tend to increase to somewhere in the 0.90 to 0.975 range. Furthermore, the worst-case increase in risk (at the least-favorable shrinkage extent) for random-coefficients tends to be only about one-third of that in the corresponding fixed-coefficient situation.

> The almost-drastic-shrinkage suggestion of Hoerl and Kennard(1970a,b) and Hemmerle(1975) again performs quite well when drastic shrinkage is appropriate (δ^{MSE} is nearly zero), reducing mean-squared-error risk by as much as 86%. But this same tactic can also increase risk by as much as 57% when almost-drastic shrinkage is inappropriate (δ^{MSE} approximately 0.90 to 0.925.)
>
> The minimum-estimated-risk suggestion [like that of Mallows(1973)] again performs well when drastic shrinkage is appropriate, reducing mean-squared-error risk by as much as 67%. But this tactic can also increase risk by as much as 17% when δ^{MSE} is approximately 0.925 to 0.95.
>
> The normal-theory, maximum-likelihood approach of Obenchain(1975,1981,1984) shrinks least aggressively even when drastic shrinkage is appropriate, reducing mean-squared-error risk by only 47% to 53%. But this tactic also never increases risk by more than 3% to 6% ...even in the least favorable situation of δ^{MSE} approximately 0.975.

Thus all three random coefficient approaches have the desirable property that they can result in a larger percentage-wise decrease in risk than their own worst-case increase in risk. And maximum likelihood limits its worst-case increase in risk to only about 6%.

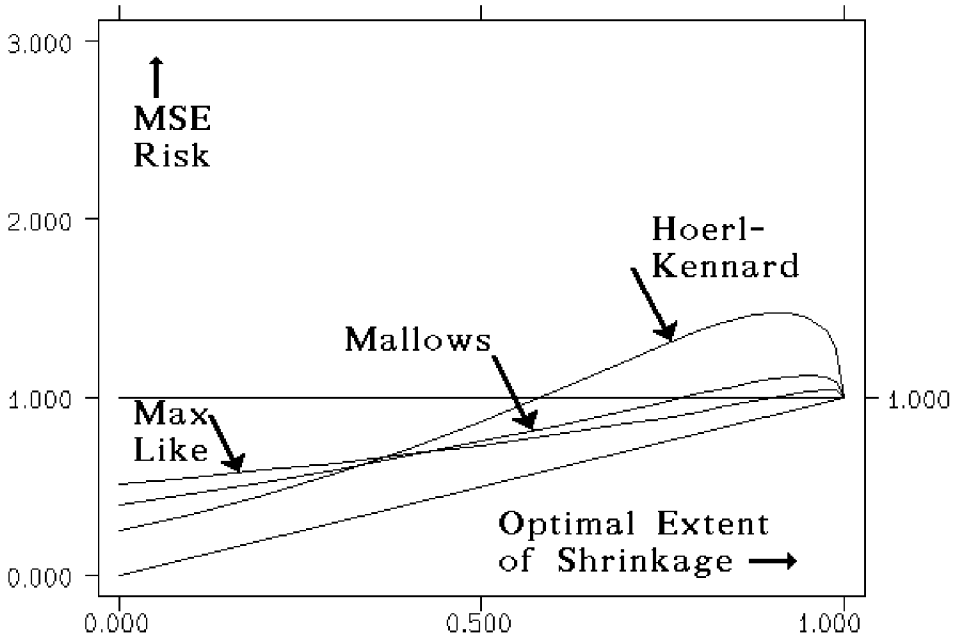
Figure 6.5 Simulated Random Coefficient Risk when Error Degrees-of-Freedom = 1.



Simulated Random Coefficient Risks for Error Degrees-of-Freedom = 1 :

delta	H-K-H	Mallows	maxLike	minMSE
0.0000	0.4577	0.5463	0.6248	0.0000
0.1000	0.5261	0.5936	0.6540	0.1000
0.2000	0.6030	0.6450	0.6856	0.1999
0.3000	0.6906	0.7012	0.7199	0.2999
0.4000	0.7914	0.7626	0.7574	0.3998
0.5000	0.9084	0.8298	0.7988	0.4997
0.6000	1.0451	0.9037	0.8450	0.5997
0.7000	1.2043	0.9845	0.8971	0.6996
0.8000	1.3844	1.0705	0.9565	0.7996
0.9000	1.5502	1.1506	1.0231	0.8997
0.9900	1.3590	1.1233	1.0585	0.9899

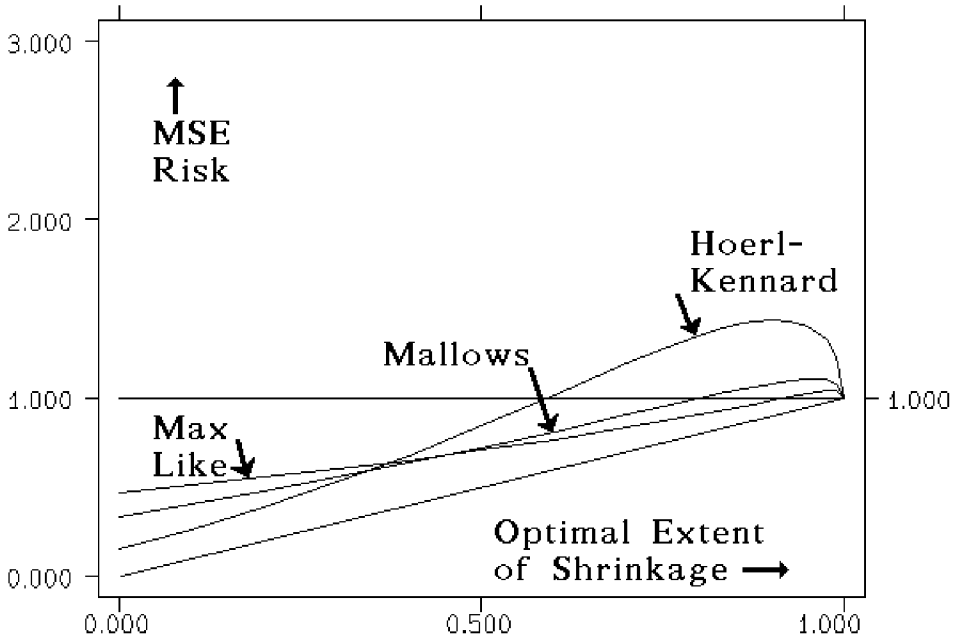
Figure 6.6 Simulated Random Coefficient Risk when Error Degrees-of-Freedom = 5.



Simulated Random Coefficient Risks for Error Degrees-of-Freedom = 5 :

delta	H-K-H	Mallows	maxLike	minMSE
0.0000	0.2464	0.3937	0.5098	0.0000
0.1000	0.3437	0.4564	0.5475	0.1001
0.2000	0.4525	0.5234	0.5881	0.2002
0.3000	0.5745	0.5949	0.6319	0.3002
0.4000	0.7110	0.6713	0.6793	0.4003
0.5000	0.8623	0.7526	0.7310	0.5003
0.6000	1.0279	0.8388	0.7878	0.6003
0.7000	1.2026	0.9290	0.8507	0.7003
0.8000	1.3705	1.0202	0.9204	0.8003
0.9000	1.4783	1.1003	0.9959	0.9002
0.9900	1.2635	1.0894	1.0418	0.9901

Figure 6.7 Simulated Random Coefficient Risk when the Variance is Known.



Simulated Random Coefficient Risks for Known Variance (Degrees-of-Freedom = ∞) :

delta	H-K-H	Mallows	maxLike	minMSE
0.0000	0.1532	0.3335	0.4672	0.0000
0.1000	0.2642	0.4023	0.5080	0.1000
0.2000	0.3882	0.4753	0.5517	0.2000
0.3000	0.5262	0.5526	0.5988	0.3000
0.4000	0.6782	0.6343	0.6497	0.4000
0.5000	0.8432	0.7205	0.7051	0.5000
0.6000	1.0173	0.8109	0.7657	0.6001
0.7000	1.1922	0.9043	0.8323	0.7001
0.8000	1.3493	0.9974	0.9056	0.8001
0.9000	1.4377	1.0797	0.9850	0.9001
0.9900	1.2318	1.0782	1.0361	0.9900

The following table, Table 6.2, again shows that the unbiased and correct-range estimates of random-coefficient variance-ratios, $\phi^2 = \sigma_\gamma^2 / \sigma^2$, have uniformly smaller mean-squared-error risk than does the maximum likelihood (F-ratio) estimate.

Table 6.2 Simulated MSE Risk in Variance-Ratio Estimation when Coefficients are Random.

First Row Label: M => F-ratio (asymptotic maximum likelihood)
U => Unbiased modification of M estimate
C = Correct range modification of U estimate

Second Row Label: Degrees-of-Freedom for Error (noise) estimation
First Column Label: MSE Optimal Shrinkage Factor, $\phi^2 / (\phi^2 + 1)$
Second Column Label: Variance Ratio, $\phi^2 = \sigma_\gamma^2 / \sigma^2$

		0.0000	0.2000	0.4000	0.6000	0.8000	0.9900
		0.0000	0.5000	0.8167	1.2250	2.0000	9.9500
M	5	29.061	38.025	61.092	125.62	457.21	163573
U	5	9.461	12.461	20.255	41.969	152.82	54130.5
C	5	8.999	11.823	19.470	41.099	152.01	54130.2
M	14	4.901	6.830	10.288	18.137	49.152	9542.0
U	14	2.600	3.943	6.343	11.764	33.008	6386.6
C	14	2.185	3.373	5.655	11.023	32.340	6386.4
M	29	3.737	5.094	7.433	12.412	29.790	3406.3
U	29	2.238	3.380	5.344	9.521	24.049	2799.5
C	29	1.833	2.825	4.677	8.805	23.408	2799.4
M	99	3.191	4.283	6.117	9.855	21.622	1089.1
U	99	2.065	3.102	4.846	8.397	19.575	1028.4
C	99	1.665	2.556	4.191	7.698	18.949	1028.3
M	∞	3.002	4.005	5.673	9.008	19.01	399.00
U	∞	2.002	3.005	4.674	8.009	18.01	398.01
C	∞	1.605	2.462	4.023	7.315	17.39	397.84

6.4.3 Summary of Risk Simulation Results

The simulated MSE risks for estimation of ϕ^2 in Tables 6.1 and 6.2 are less accurate than the 3 digit agreement achieved in the listings supporting Figures 6.2 to 6.7. After all, the first columns of Tables 6.1 and 6.2 (which give results for $\delta^{\text{MSE}} = \phi^2 = 0$) should again be identical because there is no difference between fixed-coefficients and random-coefficients in this limiting special-case. Yet we see a difference of 4.31 in simulated MSE risk when the degrees-of-freedom for error are 5, and several differences of about 0.01 when the degrees-of-freedom for error are 14 or more.

Furthermore, in both the fixed-coefficient simulation of Table 6.1 and the random coefficient results of Table 6.2, the "bad news" is that the superiority of the correct-range estimate of ϕ^2 does not necessarily translate into superior estimates of γ_i of the form $\widehat{\delta}_i \cdot c_i$ when $\widehat{\delta}_i = \widehat{\phi}_i^2 / (\widehat{\phi}_i^2 + 1)$. The correct-range approaches yield shrinkage estimates of γ_i that I feel are out-performed by the correspondingly more conservative, asymptotic maximum likelihood (F-ratio) estimates of $\widehat{\phi}_i$ in $\widehat{\delta}_i = \widehat{\phi}_i^2 / (\widehat{\phi}_i^2 + 1)$.

References for Chapter Six

Berger, J. O. (1980a). **Statistical Decision Theory: Foundations, Concepts, and Methods**. New York: Springer-Verlag.

Draper, N. R. and Van Nostrand, R. C. (1977a). "Ridge regression and James-Stein estimation: review and comments." **Technometrics** 21, 451-466.

Draper, N. R. and Van Nostrand, R. C. (1977b). "Ridge regression: is it worthwhile?" Technical Report No. 501, Department of Statistics, University of Wisconsin.

Efron, B. and Morris, C. N. (1976). "Families of minimax estimators of the mean of a multivariate normal distribution." **The Annals of Statistics** 4, 11-21.

James, W. and Stein, C. (1961). "Estimation with quadratic loss." **Proceedings of the Fourth Berkeley Symposium** 1, 361-379. University of California Press.

Jennrich, R. I. and Oman, S. D. (1986). "How much does Stein estimation help in multiple linear regression?" **Technometrics** 28, 113-121.

Johnson, N. L. and Kotz, S. (1970). **Distributions in Statistics: Continuous Univariate Distributions-2.** (Chapter 30, Noncentral F Distribution.) New York: John Wiley.

Kadiyala, K. (1980). "Some finite sample properties of generalized ridge estimators." **The Canadian Journal of Statistics** 8, 47-58.

Lindley, D. V. (1962). "Discussion." [of "Confidence sets for the mean of a multivariate normal distribution" by C. M. Stein.] **Journal of the Royal Statistical Association** B24, 285-287.

Mallows, C. L. (1973). "Some comments on Cp." **Technometrics** 15, 661-675.

Morris, C. N. (1977). "Parametric empirical Bayes inference: theory and applications" **Journal of the American Statistical Association** 78, 47-55. (with discussion, 55-65.)

Obenchain, R. L. (1978). "Good and optimal ridge estimators." **Annals of Statistics** 6, 1111-1121.

Rao, C. R. (1973). **Linear Statistical Inference and Its Applications, Second Edition**. New York: John Wiley and Sons.

Stein, C. (1955). "Inadmissibility of the usual estimate of the mean of a multivariate normal distribution." **Proceedings of the Third Berkeley Symposium** 1, 197-206. University of California Press.

Stein, C. (1962). "Confidence sets for the mean of a multivariate normal distribution." **Journal of the Royal Statistical Society** B24, 265-296.

Stein, C. (1973). "Estimation of the mean of a multivariate normal distribution." **Proceedings of the Prague Symposium on Asymptotic Statistics** 345-381.

Stein, C. (1981). "Estimation of the mean of a multivariate normal distribution." **The Annals of Statistics** 9, 1135-1151.

Thompson, J. R. (1968). "Some shrinkage techniques for estimating the mean." **Journal of the American Statistical Association** 63, 113-122.

Further Reading for Chapter Six

Dempster, A. P., Schatzoff, M. and Wermuth, N. (1977). "A simulation study of alternatives to ordinary least squares." **Journal American Statistical Association** 72, 77-91 (with discussion, pp. 91-106; see, especially, the discussion by Efron and Morris.)

Dwivedi, T. D., Srivastava, V. K. and Hall, R. L. (1980). "Finite sample properties of ridge estimators." **Technometrics** 22, 205-212.

Gibbons (Galarneau), D. I. (1981). "A simulation study of some ridge estimators." **Journal of the American Statistical Association** 76, 131-139.

Goto, M. (1979). "Choice of shrinkage factors in the generalized ridge regression." **Math Japonica** 24, 153-173.

Goto, M. and Matsubara, Y. (1979). "Evaluation of ordinary ridge regression." **Bulletin of Mathematical Statistics,** Research Association of Statistical Sciences, 19, 1-35.

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

Hemmerle, W. J. (1975). "An explicit solution for generalized ridge regression." **Technometrics**, 17, 309-314.

Hemmerle, W. J. and Brantley, T. F. (1978). "Explicit and constrained generalized ridge estimation." **Technometrics** 20, 109-119.

Hemmerle, W. J. and Carey, M. B. (1981). "Some properties of generalized ridge estimators." Department of Computer Science and Experimental Statistics, University of Rhode Island.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression." **Biometrics** 32, 1-49.

Hoerl, A. E. and Kennard, R. W. (1970a,b). "Ridge regression: biased estimation for nonorthogonal problems." and "Ridge regression: applications to nonorthogonal problems." **Technometrics** 12, 55-67 and 69-82.

Hoerl, A. E., Kennard, R. W. and Baldwin, K. F. (1975). "Ridge regression: some simulations." **Communications in Statistics** A4, 105-123.

Hoerl, R. W., Schuenemeyer, J. H. and Hoerl, A. E. (1986). "A simulation of biased estimation and subset selection regression techniques." **Technometrics** 28, 369-380.

Krishnamurthi, L. and Rangaswamy, A. (1987). "The equity estimator for marketing research." **Marketing Science** 6, 336-357.

Krishnamurthi, L. and Rangaswamy, A. (1990). "Response function estimation using the equity estimator." (Revised.) Working Paper No. 89-030R. Philadelphia: The Wharton School, University of Pennsylvania.

Lawless, J. F. (1975). "A note on certain types of regression estimators and their mean squared error of prediction properties." University of Waterloo, Canada.

Lawless, J. F. and Wang, P. (1976). "A simulation study of ridge and other regression estimators." **Communications in Statistics**, 5, 307-323.

McDonald, G. C. and Galarneau, D. I. (1975). "A monte carlo evaluation of some ridge-type estimators." **Journal of the American Statistical Association**, 70, 407-416.

Newhouse, J. P. and Oman, S. D. (1971). "An evaluation of ridge estimators." Rand Report No. R-716-PR (28 pages.) Santa Monica, California; The Rand Corporation.

Obenchain, R. L. (1975b). "Ridge analysis following a preliminary test of the shrunken hypothesis." **Technometrics**, 17, 431-441. (Discussion: McDonald, G. C., 443-445.)

Obenchain, R. L. (1976). "Methods of ridge regression." **Proceedings of the Ninth International Biometric Conference**, Invited Papers, Volume One, 37-57, Boston.

Obenchain, R. L. (1981). "Maximum likelihood ridge regression and the shrinkage pattern hypotheses." Abstract 81t-23. **I.M.S. Bulletin** 10, 37.

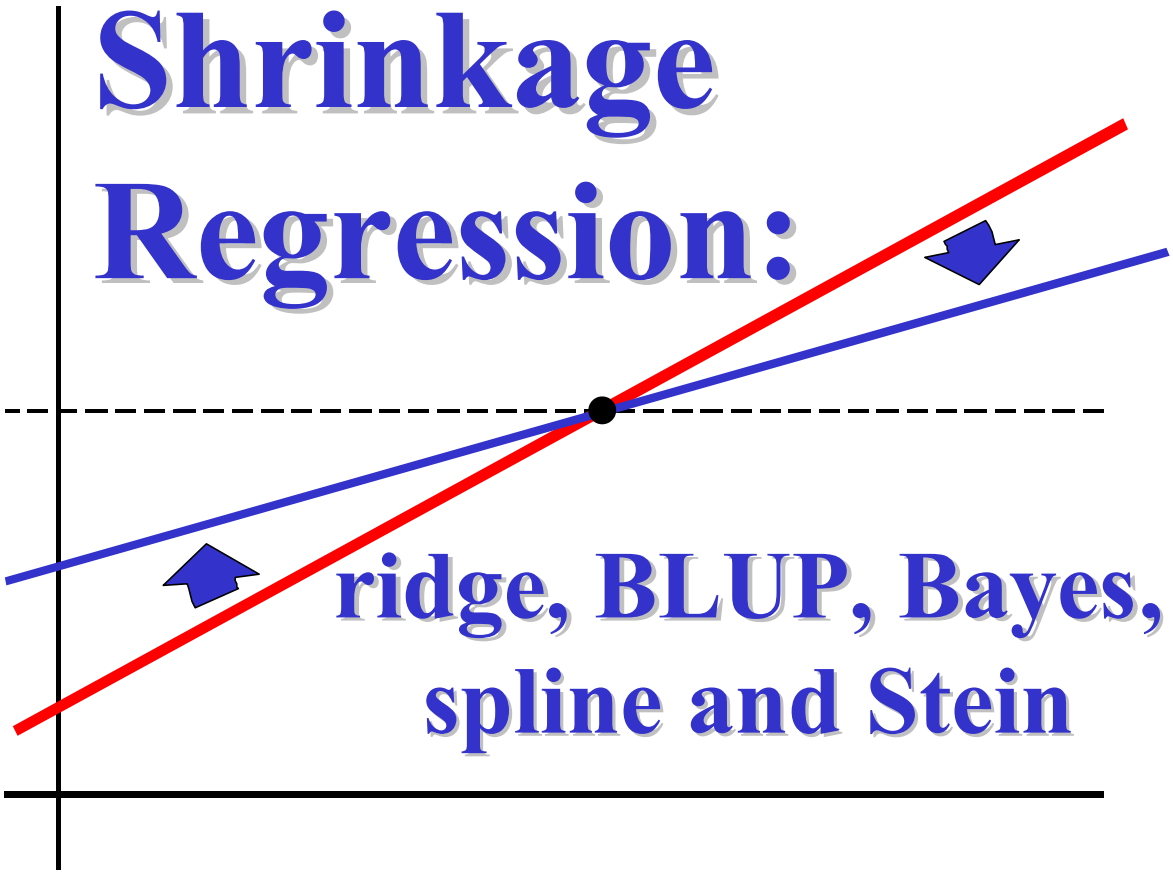
Obenchain, R. L. (1984). "Maximum likelihood ridge displays." (Proceedings of the Fordham Ridge Symposium, ed. H. D. Vinod.) **Communications in Statistics** A13, 227-240.

Sclove, S. (1968). "Improved estimators of coefficients in linear regression." **Journal of the American Statistical Association** 63, 596-606.

Trenkler, G. and Trenkler, D. (1981). "Estimable functions and reduction of mean squared error." **Methods of Operations Research** 44, 225-234. Oelgeschlager, Gunn & Hain, Cambridge, Mass.

Wichern, D. W. and Churchill, G. A. (1978). "A comparison of ridge estimators." **Technometrics** 20, 301-311.

Yancey, T. A. and Judge, G. G. (1977). "A Monte Carlo comparison of traditional and Stein-rule estimators under squared error loss." **Journal of Econometrics** 4, 285-294.



Chapter 08: Bayesian Formulations

Bob Obenchain, Ph.D.
softR$_X$ freeware
13212 Griffin Run
Carmel, Indiana 46033-8835

Chapter 8: BAYESIAN FORMULATIONS

Here in Chapter 8 we discuss Bayesian methods, both hierarchical and empirical, for defining the form and extent of shrinkage of sample estimates towards a subjective prior distribution. We find some striking parallels between Bayesian and classical shrinkage methodologies; formulas for point estimates of regression coefficients can frequently be made to agree exactly. But we also find profound differences; the Bayesian posterior variance of a point estimate is larger than the classical variance of that same estimate. We discuss not only why this type of difference exists but also point out some specific implications for statistical inference.

8.1 Bayesian Conjugate-Normal Linear-Model Formulations

Lindley and Smith(1972) describe a Bayesian formalism for hierarchical (multi-stage) analyses of linear models using conjugate multivariate-normal prior distributions. This formalism expresses unknown parameters at each stage of an analysis in terms of a linear model at the previous, lower stage. But, although dispersion matrices at each stage can be arbitrary, they must be known. And, at the final stage, both the mean vector and the dispersion matrix must be known. Here, we will be more interested in very simple, 2-stage analyses than in 3-or-more-stage (priors-on-priors) models. Thus our discussion will only rarely dwell deeper into Bayes' theory/practice than what is provided by "classic" reference works such as those of Raiffa and Schlaifer(1961) and Box and Tiao(1973).

The notation $y \sim N(\mu , D)$ will mean here that the column vector y has a multivariate normal distribution with mean given by the column vector μ and variance-covariance matrix given by the positive semi-definite matrix D. Similarly, $y \mid \theta \sim$ will mean that the conditional distribution of y given θ is being defined. In this notation, the fundamental lemma of Lindley and Smith(1972), pages 4-5, states that:

> LEMMA: If the sampling distribution of the response, y, is $y \mid \theta_1 \sim N(A_1 \theta_1 , D_1)$, where θ_1 is a $p_1 \times 1$ parameter vector, and the prior distribution is $\theta_1 \mid \theta_2 \sim N(A_2 \theta_2 , D_2)$ where θ_2 is a $p_2 \times 1$ parameter vector, then the marginal (unconditional) distribution of y is
>
> $$y \sim N(A_1 A_2 \theta_2 , D_1 + A_1 D_2 A_1^T) \qquad \{ 8.1 \}$$
>
> and the posterior (conditional) distribution of θ_1 given y is
>
> $$\theta_1 \mid y \sim N(B b , B) \qquad \{ 8.2 \}$$

where $B^{-1} = A_1^T D_1^{-1} A_1 + D_2^{-1}$ and $b = A_1^T D_1^{-1} y + D_2^{-1} A_2 \theta_2$.

To apply this lemma and demonstrate that a simple 2-stage Bayesian formalism produces generalized shrinkage estimators, we first make the identifications $A_1 \theta_1 = X \beta$ and $D_1 = \sigma^2 \cdot I$. Thus we are using a "point prior" on the error variance (i.e. proceeding as if σ^2 where known) and $B^{-1} = \sigma^{-2} \cdot X^T X + D_2^{-1}$. Next, we set the prior mean value for β to ZERO by taking $\theta_2 = 0$ and assure that D_2^{-1} (and D_2) will be simultaneously diagonalizable with $X^T X$ by restricting attention to prior variance-covariance matrices of the general form $D_2 = \sigma^2 \cdot G K^{-1} G^T$, where K is a diagonal R × R matrix and G is the P × R semi-orthogonal matrix of direction cosines for the principal axes of X , as in equation { 2.8 }. Now the Bayes point estimate is the mean, B b , of the posterior distribution of β given y (as well as given X) and this mean vector is of the general form:

$$E (\beta | y , X) = G (\Lambda+K)^{-1} \Lambda G^T y = G \Delta c = b^{\bigstar} \qquad \{ 8.3 \}$$

as in { 3.1 } , where $\Delta = \Lambda (\Lambda + K)^{-1} = K^{-1} (\Lambda^{-1} + K^{-1})^{-1}$ is the diagonal matrix of generalized shrinkage factors and c is the vector of uncorrelated components of the least-squares estimator. In other words, we have now demonstrated that all generalized shrinkage estimators are 2-stage Bayes estimates. This includes, of course, the special case of shrinkage estimates in the 2-parameter shrinkage family of { 3.9 } where $K = k \cdot \Lambda^Q$, the scalar Q determines the shape of the shrinkage path, and the scalar k (or, equivalently, $MCAL = R - \delta_1 - \ ... - \delta_R$) determines the extent of shrinkage. [Bayes estimates of more general form than { 8.3 } can, of course, result from choices of $\theta_2 \neq 0$ and/or $D_2 \neq \sigma^2 \cdot G K^{-1} G^T$.]

The above observation goes a long way, perhaps, towards explaining why many people apparently think that shrinkage estimation methods are Bayesian. But, wait a second! The Bayesian variance-covariance matrix, B , of the posterior distribution of the regression coefficient vector, β , given the observed vector of responses, y , (as well as given X) is of the general form:

$$V (\beta | y , X) = \sigma^2 \cdot (X^T X + G K G^T)^{-1} = \sigma^2 \cdot G \Delta \Lambda^{-1} G^T . \qquad \{ 8.4 \}$$

Note that the implied Bayesian dispersion (variance, co-variance) matrix for $G^T \beta = \gamma$ is the diagonal matrix $\sigma^2 \Delta \Lambda^{-1}$ and that this matrix has the same functional form as the classical, fixed-effect minimum risk, { 4.7 }, in $G^T b^{\bigstar} = \Delta \gamma$, which is achieved only when $\Delta = \Delta^{MSE}$. Note, in particular, the Bayesian dispersion is usually larger than the classical dispersion of equation { 3.4 },

$$V (b^{\bigstar} | X) = \sigma^2 \cdot G \Delta^2 \Lambda^{-1} G^T .$$

After all, $\Delta^2 \leq \Delta$ when shrinkage factors are restricted to their "usual" range of $0 \leq \delta_i < 1$ for $1 \leq i \leq R$. In fact, strict in-equality ($\Delta^2 < \Delta$) holds whenever none of the shrinkage factors

is an actual ZERO. As we stress below, this difference in dispersion matrices has profound, practical implications for statistical inference.

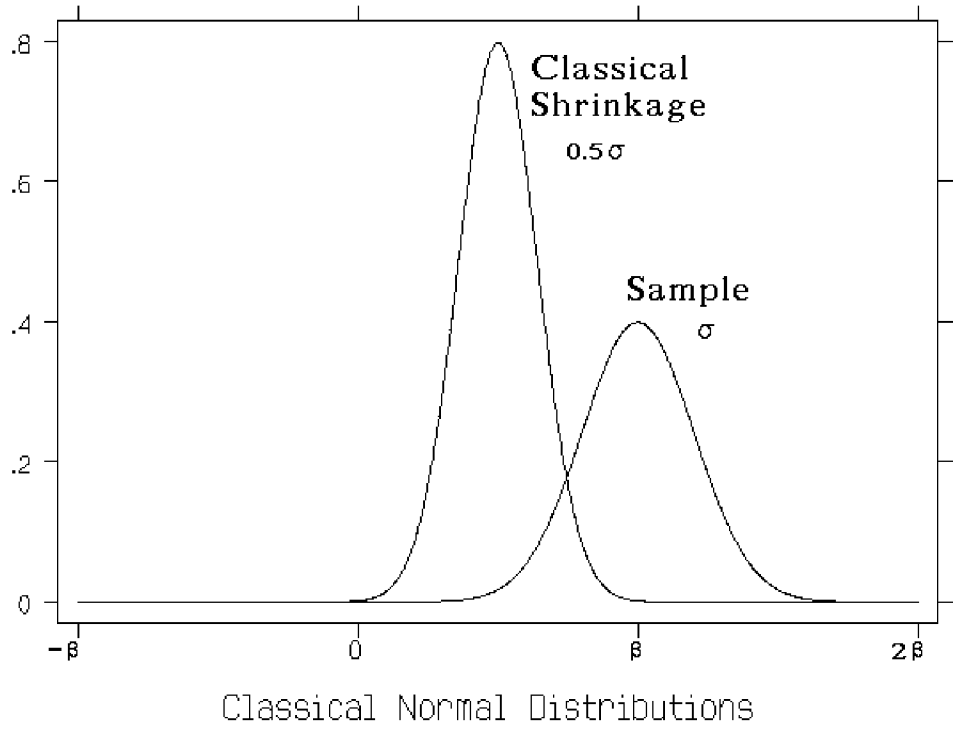
> Before starting our arguments on why Bayesian variances exceed their classical counterparts, we comment that substitution of estimates for unknown parameters into formulas { 8.3 } and { 8.4 } is commonly known as the "naive" empirical Bayes approach. This approach is apparently said to be naive because, relative to a full-blown hierarchical Bayes analysis, variances are thereby under estimated! For example, Ghosh(1992), pages 153-154, discusses an analysis that illustrates how naive substitutions ignore the "uncertainty involved in estimating the prior parameters when estimating the posterior variance." Then Ghosh(1992) argues [pages 168-173] that, although the empirical Bayes method of Morris(1983) is "an attempt to approximate a bonafide hierarchical Bayes procedure, and is clearly superior to a naive empirical Bayes procedure", variances are then over estimated by 11% in one example (and might be as much as 30% too large.) All that I really wish to stress here is that equation { 8.4 } apparently represents some sort of lower-bound for the variance of the shrinkage estimator { 8.3 } from Bayesian points-of-view. And yet this minimum Bayesian variance can still be considerably larger that the classical variance of that same estimator. Here's why...

Bayesian estimators incorporate added information from the prior distribution into the analysis. In fact, Bayes estimates are considered to be unbiased relative to combined sample and prior information about β. In other words, the variance-covariance matrix of a Bayes estimate is also its mean-squared-error matrix! In particular, the rank 1 squared-biases matrix of the classical formulation, the $(I - \Delta)\gamma\gamma^{T}(I - \Delta)$ term in { 4.2 }, is absent from the Bayesian formulation. And every choice for Δ yields Bayes risks for true components, γ, that behave like the minimum classical risks in Δ c achieved only at $\Delta = \Delta^{MSE}$.

From a Bayesian point-of-view, the more drastic is the shrinkage (the smaller is the Δ) imposed by a highly "informative" prior, the better-off one ends-up being! In other words, conflict between prior and sample information can be tolerated because "shrinkage" will effect a compromise. The more distinct/remote is the prior distribution from the sampling distribution, the more distinct/remote will be the posterior estimate from the sample estimate. In fact, one's prior distribution is more informative in these large-separation cases, and the Bayes posterior estimate ends up being correspondingly more precise.

Classical fixed-effect analyses of shrinkage estimators assume that bias is being introduced into the analysis. The multiplicative δ-factors in classical shrinkage formulas enter variance formulas as δ^2-factors. In other words, the standard deviations (square roots of variances) of classical shrinkage estimators are multiplied by the same δ-factors as are expected values; expected values and standard deviations thus change at exactly the same rate in classical shrinkage analyses. This point is illustrated in Figure 8.1 where a shrinkage factor of δ=0.5 changes the expected value of an estimate from β to $\beta/2$ and its standard deviation from σ to $\sigma/2$.

Figure 8.1 The Classical Shrinkage Formulation



Classical Normal Distributions

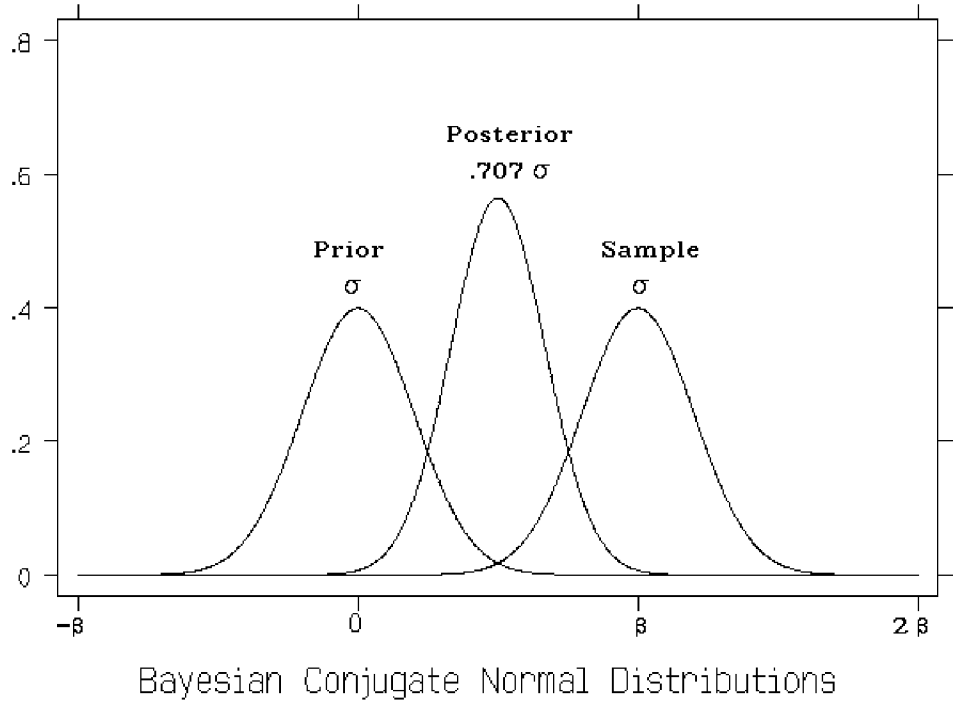
Shrinkage has no effect whatsoever on classical fixed-effect statistical inferences for regression coefficients that are based upon F-ratios and t-statistics. After all, a t-statistic is a ratio with an unbiased estimate of the numerical size of an effect in its numerator and a root-mean-square estimate of the corresponding standard deviation in its denominator; a F-ratio is the square of a t-statistic (or a sum-of-squares of several t-statistics with the same denominator). Anyway, there are convincing arguments [see Obenchain(1977) for details] that these ratios are actually invariant under classical shrinkage. In other words, classical fixed-effect shrinkage does not produce confidence intervals/regions for regression coefficients that are shifted in location and/or differnet in size from those derived by ordinary least-squares theory.

ASIDE: One might consider forming a confidence set for δ times β. Relative to a classical confidence set for β, the corresponding classical confidence set for $\delta \cdot \beta$ would be shifted in location (being centered at $\delta \cdot b^{o}$) and would be smaller in size (being based on dispersion $\delta \cdot s$) whenever $0 \leq \delta < 1$. On the other hand, confidence sets for $\delta \cdot \beta$ are of relatively little practical interest compared with the confidence set for the full β vector!

In summary, classical fixed-effect shrinkage methods are best applied on a contingency basis. The data at hand may provide convincing evidence of reduced dispersion that will more than offset the introduction of squared-bias, yielding an overall reduction in mean-squared-error. However, although one has the option of shrinking classical point estimates of regression coefficients, their corresponding fixed-effect set estimates remain unchanged. On the other hand, point and set estimates usually would change or shift, at least a little, if fixed-effects in a classical model were declared random. After all, BLUEs are then replaced with shrunken

BLUPs and confidence intervals/regions are then constructed using variance-component estimates!

Figure 8.2 A Bayesian Shrinkage Formulation



Because Bayesian posterior variances decrease in direct proportion to their δ shrinkage factors, Bayesian standard deviations decrease at a slower ($\delta^{1/2}$) rate than do their mean values. This point is illustrated in Figure 8.2, above, where a Bayes shrinkage factor of δ=0.5 results because the sample distribution (centered at β) and the prior distribution (centered at 0) are of exactly equal precision, σ. This Bayesian shrinkage produces a posterior distribution with expected value $\beta/2$, but the posterior standard deviation is $\sigma/\sqrt{2} = 0.707\ \sigma$ rather than $\sigma/2$.

Bayesian formulas for posterior F-ratios and t-statistics that measure differences between a posterior estimate and its prior mean tend to be “shrunken" in the sense that their numerators (effect sizes) have decreased more than their denominators (uncertainty measures.) This, of course, weakens any evidence that the posterior estimate might be discordant with the prior mean. In fact, Bayesian highest-posterior-density intervals/regions resulting from an informative prior for regression coefficients definitely are shifted in location (towards the prior) and are smaller in size than are the corresponding classical (frequentist) intervals/regions. By incorporating added information from the prior into the analysis, Bayes procedures end up “shrinking" highest-posterior-density set estimates as well as point estimates of regression coefficients.

8.2 Bayesian Diagnostic Checking

Informal methods for determining the effects of changes in one's Bayesian prior distribution upon the implied posterior distribution are commonly called "sensitivity analyses," Winkler(1972). Modern hardware/software reduces the implied computational burden to the point where at least some sort of diagnostic checking would seem to be a mandatory component of even in the most routine of Bayesian analyses. Box and Tiao(1973) and Berger(1980a, 1980b, 1983) suggest some more formal methods under the general title of Bayesian "robustness." And uncertainty about one's prior apparently motivates the 3rd-and-higher stages in the hierarchical approach of Lindley and Smith(1972).

In his "model adequacy" approach to "assessing the prior" that yields shrinkage regression estimates [Box(1980), Section §3], Box considers "predictive checks" derived using the marginal distribution, { 8.1 }. [This marginal distribution can be called "predictive" in the sense that it describes the expected behavior of sample data for the current analysis, but this marginal distribution is definitely distinct from the "predictive" distribution of a future sample, Zellner and Chetty(1965) or Aitchison and Dunsmore(1975), that results from integrating the sampling distribution over the posterior distribution.] In any case, we note that the mean and variance of the marginal distribution are:

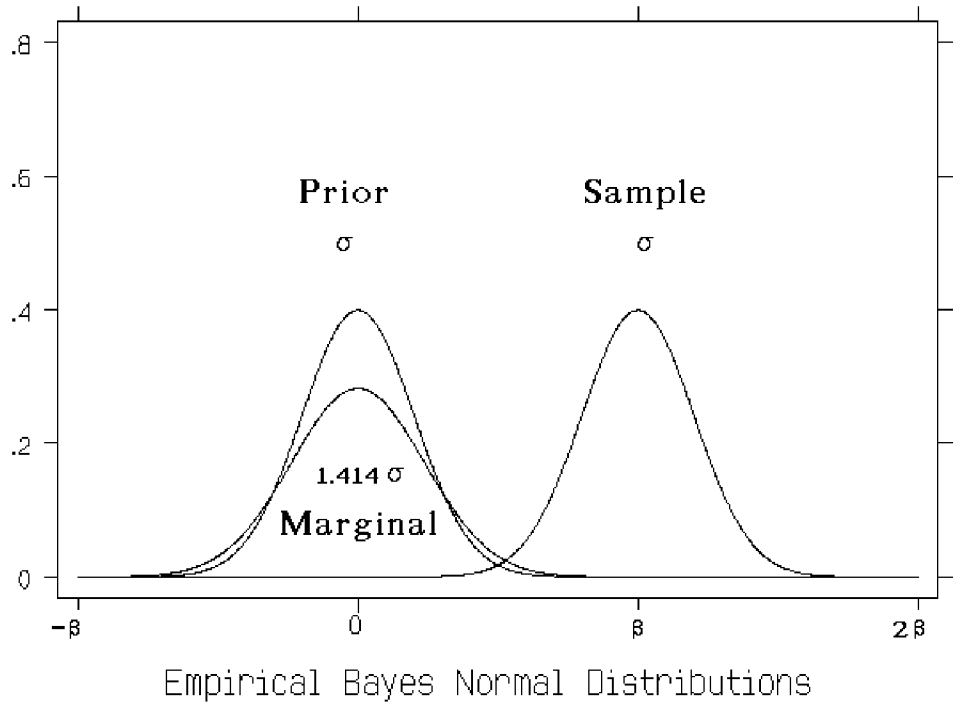
$$E(\beta \mid X) = 0 \quad \text{(the prior mean)} \qquad \{ 8.5 \}$$

and

$$V(\beta \mid X) = \sigma^2 \cdot G(I - \Delta)^{-1} \Lambda^{-1} G^T . \qquad \{ 8.6 \}$$

Thus, relative to the distribution of sample estimates, the marginal distribution is not only shifted in location (to the point that it totally ignores sample information!) but also has increased dispersion. These points are illustrated in Figure 8.3 where a Bayes shrinkage factor of δ=0.5 again results because the sample distribution (centered at β) and the prior distribution (centered at 0) are of exactly equal precision, σ. Note that the marginal distribution also has an increased standard deviation of $\sqrt{2}\,\sigma = 1.414\,\sigma$.

Figure 8.3 A Bayesian Marginal Distribution



Now Box(1980), equations (3.1) to (3.10), points out that his “predictive check" agrees with the Theil(1963) measure of the “compatibility of prior and sample information" and is defined as follows: For the extent of shrinkage implied by a given set of factors, Δ , calculate the F-ratio that measures the squared-distance between the least-squares estimates vector and the marginal mean in the metric of the marginal dispersion, namely

$$\text{Bayes predictive F-ratio} \;=\; c^T (I - \Delta) \Lambda c / (s^2 R), \qquad \{ 8.7 \}$$

where s^2 is the sample residual-mean-square. Then calculate the observed significance level of this F-ratio, which is the probability that a random variate with a central F-distribution (with R numerator degrees-of-freedom and N – R – 1 denominator degrees-of-freedom) would exceed that observed F value. This observed significance level allows any choice for the extent of shrinkage, Δ , to be “criticized."

It seems to me, at least, that the corresponding classical statistic would measure the squared-distance between the least-squares estimates and the shrunken estimates in the metric of the sampling dispersion, namely

$$\text{Classical F-ratio} \;=\; c^T (I - \Delta)^2 \Lambda c / (s^2 R). \qquad \{ 8.8 \}$$

Note that, just as Δ versus Δ^2 provide the distinction between the Bayes and classical dispersion matrices of { 8.4 } and { 3.4 }, $(I - \Delta)$ versus $(I - \Delta)^2$ provide the distinction between the Bayes and classical statistics of equations { 8.7 } and { 8.8 }. By the way, Obenchain(1977) called the observed significance level of { 8.8 } the associated probability of classical ridge shrinkage, while McCabe(1978) termed this same quantity the α-acceptability for that extent of shrinkage. Also, note that all significance levels (classical and Bayesian) are being computed relative to the same central F-distribution (with R numerator degrees-of-freedom and $N - R - 1$ denominator degrees-of-freedom.)

Next, note that there is a consistent difference between the Bayesian and classical observed significance levels associated with a given extent of shrinkage. Because the classical F-ratio of { 8.8 } is almost always smaller, numerically, than is the Bayes' predictive F-ratio of { 8.7 }, its significance level is almost always larger (less significant) than that given by the Bayesian evaluation of the same shrinkage. In other words, once a Bayesian becomes "introspective" about his/her specific choice of location and/or spread for a prior distribution, he/she is almost always more critical of his/her own choice than a classicist would be when evaluating the exact same form and extent of shrinkage.

8.3 More Bayes' Measures of the Extent of Shrinkage

There are at least two ways to quantify the extent of shrinkage employed in a given set of Bayes estimates for linear model coefficients.

Theil(1963) describes a variety of ideas, many of which were expanded on by later authors. For example, Theil(1963) proposes the "f-class" of mixed (empirical Bayes) estimators for regression, which provide a form of generalized shrinkage towards an origin space, and suggests (page 404) that these estimates be plotted as in a ridge TRACE type of display. And, as observed earlier, Theil(1963), equation (3.3), describes a special case of equation { 8.7 }. However, in my opinion, the primary contribution of Theil(1963) is his demonstration of uniqueness properties of a certain Bayesian measure of extent of shrinkage originally introduced by Schlaifer.

Theil started by asking a question like... "What proportions of posterior relative precision in a Bayes estimate are due, respectively, to sample information and to prior information." This is like asking... "What are the relative contributions of matrices A and B to the matrix $(A + B)^{-1}$?" Specifically, suppose that a function $g(A , B)$ is to be our measure the contribution of A to $(A + B)^{-1}$. Theil(1963) argued that the following four requirements on $g(A , B)$ seem reasonable.

(i) Adding-Up Criterion: $g(A , B) + g(B , A) \equiv 1$.

(ii) Zero Unit Criterion: $g(0 , B) = 0$ when $B \neq 0$

and

$g(A , 0) = 1$ when $A \neq 0$.

(iii) Invariance Under Nonsingular Linear Transformations, K , of Predictors:

$$g(K^T A K , K^T B K) \equiv g(A , B) .$$

(iv) Linearity Criterion: If A_1 , B_1 , A_2 and B_2 are such that $A_1 + B_1 = A_2 + B_2$ and p and q are two non-negative scalars that sum to one, then

$$g(p \cdot A_1 + q \cdot A_2 , p \cdot B_1 + q \cdot B_2) \equiv p \cdot g(A_1 , B_1) + q \cdot g(A_2 , B_2) .$$

Then Theil(1963) demonstrated that the unique measure satisfying all four of the above criteria is

$$g(A , B) = \text{trace}[A (A + B)^{-1}] / R , \qquad \{ 8.9 \}$$

when A and B are R × R matrices. Applying this result to the Bayes posterior variance, { 8.4 }, where the sampling precision is $A = \sigma^{-2} \cdot G \Lambda G^T$ and the prior precision is $B = \sigma^{-2} \cdot G K G^T$, we find that the proportion of posterior precision due to sample information is

$$g(\Lambda , K) = \text{trace}[\Lambda (\Lambda + K)^{-1}] / R = \sum \delta_i / R , \qquad \{ 8.10 \}$$

which is (R − MCAL) / R. In other words, when the multicollinearity allowance is MCAL = 0 [so that $\Delta = I$, and no shrinkage gets applied], then all posterior precision derives from sample information. But, at the other extreme of MCAL = R [where $\Delta = 0$, and total shrinkage to the prior mean is enforced], then none of posterior precision is derived from sample information.

Lindley(1980) writes that the "only satisfactory inference definition of information is surely Shannon's" and gives a formula that can be derived as follows: The Bayes estimator (posterior mean) of equation { 8.3 } can be written as a linear transformation, $b^{\bigstar} = (I + Z)^{-1} b^o$, of the sample (least-squares) estimator, b^o, where $Z = G \Lambda^{-1} K G^T$. Now Shannon's measure of information gain, posterior minus prior, is given by the corresponding difference in the expected values of the log likelihoods. For an R-dimensional multivariate normal distribution with dispersion matrix Σ , the expected log likelihood is $E(\ln L) = -\frac{1}{2} \cdot [R \cdot \ln(2\pi) + R + \ln |\Sigma|]$. As a result, Shannon's measure of information gain can be written as

$$\Im = \tfrac{1}{2} \cdot \ln[| I + Z | / | Z |] = -\tfrac{1}{2} \cdot \sum \ln(1 - \delta_i) . \qquad \{ 8.11 \}$$

Thus Shannon's measure of information gain is $\Im = +\infty$ when $\Delta = I$ [because the prior suggests no shrinkage whatsoever in this extreme case] and $\Im = 0$ when $\Delta = 0$ [because the posterior and prior coincide in this extreme case.]

8.4 Nonconjugate Bayes Formulations

I would like to make a couple of observations about nonconjugate Bayes analyses of linear models even though I am not sufficiently familiar with this literature to critique it here in any real detail. The nonconjugate analyses proposed by Draper and Van Nostrand(1977b) and Berger(1980b,1983) yield regression coefficient estimates of specifically nonlinear form. Their resulting risk (mean-squared-error) matrices strike me as being potentially more realistic than { 8.4 }. Unfortunately, the equations that define these sorts of estimates are sufficiently complicated, mathematically, that nothing short of detailed computational experience in applying these techniques to a wide variety of numerical examples would be adequate to appreciate how well they might perform in actual practice.

8.5 An Empirical Bayes Likelihood Approach

In his rejoinder to the discussion of his paper, Morris(1983) observes

> "Several discussants have gathered that `empirical Bayes' means plugging non-Bayesian estimates of the prior distribution into Bayes rules. I believe nothing in the empirical Bayes paradigm, or in frequency theory for that matter, forbids use of Bayes rules."

I agree and would add the thoughts: Bayes theorem is, after all, a theorem in classical statistics. Why should anybody feel hesitant to apply these tools in ways that they feel are appropriate and reasonable?

The empirical Bayes minus-two-log-likelihood factor of Efron and Morris(1977) for evaluating the extent of shrinkage is also based upon the marginal (predictive) distribution of equations { 8.1 }, { 8.5 }, and { 8.6 }. The minus-twice-log-likelihood resulting from treating the vector of least squares estimates for regression coefficients as if it were an observation from this marginal distribution [using the residual-mean-square s^2 as one's estimate of σ^2] is

$$-2 \cdot \ln(\mathrm{ML}) = R \cdot \ln(2 \cdot \pi \cdot s^2) + \sum_{i=1}^{R} \{ F_i \cdot (1 - \delta_i) - \ln[\lambda_i \cdot (1 - \delta_i)] \}, \qquad \{ 8.12 \}$$

where $F_i = c_i^2 \cdot \lambda_i / s^2$ is again the F-ratio of equation { 2.22 } for testing the statistical significance of the i-th uncorrelated component of the least-squares vector. When actually applying this criterion, Efron and Morris(1977) suggest simply ignoring all of the terms in { 8.12 } that do not change as shrinkage occurs. The factor they suggest computing is thus

$$\mathrm{EBAY} = \sum_{i=1}^{R} F_i \cdot (1 - \delta_i) + 2 \cdot \Im, \qquad \{ 8.13 \}$$

where $\Im = -\frac{1}{2} \cdot \sum \ln(1 - \delta_i)$ is Shannon's measure of information gain from equation { 8.11 }.

References for Chapter Eight

Aitchison, J. and Dunsmore, I. R. (1975). **Statistical Prediction Analysis**. Cambridge University Press.

Berger, J. O. (1980a). **Statistical Decision Theory: Foundations, Concepts, and Methods**. New York: Springer-Verlag.

Berger, J. O. (1980b). "A robust generalized Bayes estimator and confidence region for a multivariate normal mean." **Annals of Statistics** 8, 716-761.

Berger, J. O. (1983). "The robust bayesian viewpoint." **Robustness in Bayesian Statistics**. J. Kadane, ed. Amsterdam: North Holland.

Box, G. E. P. (1980). "Sampling and Bayes' inference in scientific modeling and robustness." **Journal of the Royal Statistical Society** A143, 383-404. (with discussion 404-430.)

Box, G. E. P. and Tiao, G. C. (1973). **Bayesian Inference in Statistical Analysis.** Reading, Massachusetts: Addison-Wesley.

Draper, N. R. and Van Nostrand, R. C. (1977a). "Ridge regression and James-Stein estimation: review and comments." **Technometrics** 21, 451-466.

Draper, N. R. and Van Nostrand, R. C. (1977b). "Ridge regression: is it worthwhile?" Technical Report No. 501, Department of Statistics, University of Wisconsin.

Efron B. and Morris, C. N. (1973). "Stein's estimation rule and its competitors." **Journal of the American Statistical Association** 68, 117-130.

Efron B. and Morris, C. N. (1975). "Data analysis using Stein's estimator and its generalizations." **Journal of the American Statistical Association** 70, 311-319.

Efron B. and Morris, C. N. (1977). "Comment" [on "A simulation study of alternatives to ordinary least squares," by Dempster, Schatzoff, and Wermuth.] **Journal of the American Statistical Association** 72, 91-93.

Ghosh, M. (1992). "Hierarchical and empirical Bayes multivariate estimation." **Current Issues in Statistical Inference: Essays in Honor of D. Basu**. Institute of Mathematical Statistics, LECTURE NOTES-MONOGRAPH SERIES #17, 151-177.

Lindley, D. V. and Smith, A. F. M. (1972). "Bayes estimates for the linear model." **Journal of the Royal Statistical Society** B34, 1-72.

Lindley, D. V. (1980). "Comment" [on "A critique of some ridge methods" by Smith and Campbell.] **Journal of the American Statistical Association** 75, 94-95.

McCabe, G. P. (1978). "Evaluation of regression coefficients using α-acceptability." **Technometrics** 20, 131-139.

Morris, C. N. (1983). "Parametric empirical Bayes inference: theory and applications" **Journal of the American Statistical Association** 78, 47-55. (with discussion, 55-65.)

Obenchain, R. L. (1977). "Classical F-tests and confidence regions for ridge regression." **Technometrics** 19, 429-439.

Obenchain, R. L. (1981). "Maximum likelihood ridge regression and the shrinkage pattern hypotheses." Abstract 81t-23. **Institute of Mathematical Statistics Bulletin** 10, 37.

Raiffa, H. and Schlaifer, R. (1961). **Applied Statistical Decision Theory**. Harvard University Press.

Rao, C. R. (1973). **Linear Statistical Inference and Its Applications, Second Edition**. New York: John Wiley and Sons.

Rolph, J. E. (1976). "Choosing shrinkage estimators for regression problems." **Communications in Statistics** A5, 789-802.

Theil, H. (1963). "On the use of incomplete prior information in regression analysis." **Journal of the American Statistical Association** 58, 401-414.

Thisted, R. (1976). "Ridge regression, minimax estimation, and empirical Bayes methods." **Technical Report No. 28, Division of Biostatistics**, Stanford University.

Winkler, R. L. (1972). **An Introduction to Bayesian Inference and Decision**. New York: Holt, Rinehart and Winston.

Zellner, A. and Chetty, V. K. (1965). "Prediction and decision problems in regression models from the Bayesian point of view." **Journal of the American Statistical Association** 60, 608-616.

Further Reading for Chapter Eight

Anderson, R. L. and Battiste, E. L. (1975). "The use of prior information in linear regression analysis." **Communications in Statistics** 4, 497-517.

Bacon, R. W. and Hausman, J. A. (1974). "The relationship between ridge regression and the minimum mean square error estimator of Chipman." **Oxford Bulletin of Economics and Statistics**, 36, 115-124.

Banerjee, K. S. and Carr, R. N. (1971). "A comment on ridge regression, biased estimation for nonorthogonal problems." **Technometrics** 13, 895-898.

Brown, P. and Payne C. (1975). "Election night forecasting." **Journal Royal Statistical Society** A138, 463-498.

Guilkey, D. K. and Murphy, J. L. (1975). "Directed ridge regression techniques in cases of multicollinearity." **Journal American Statistical Association** 70, 769-775.

Hsiang, T. C. (1976). "A Bayesian view on ridge regression." **The Statistician** 24, 267-268.

Goldstein, M. and Smith, A. F. M. (1974). "Ridge-type estimators for regression analysis." **Journal Royal Statistical Society** B36, 284-291.

Goldstein, M. (1976). "Bayesian analysis of regression problems." **Biometrika** 63, 51-58.

Good, I. J. (1965). **The Estimation of Probabilities, An Essay on Modern Bayesian Methods.** Cambridge, Massachusetts: M.I.T. Press.

Leamer, E. E. (1977). "Valley regression: biased estimation for orthogonal problems." Department of Economics, University of California, Los Angeles.

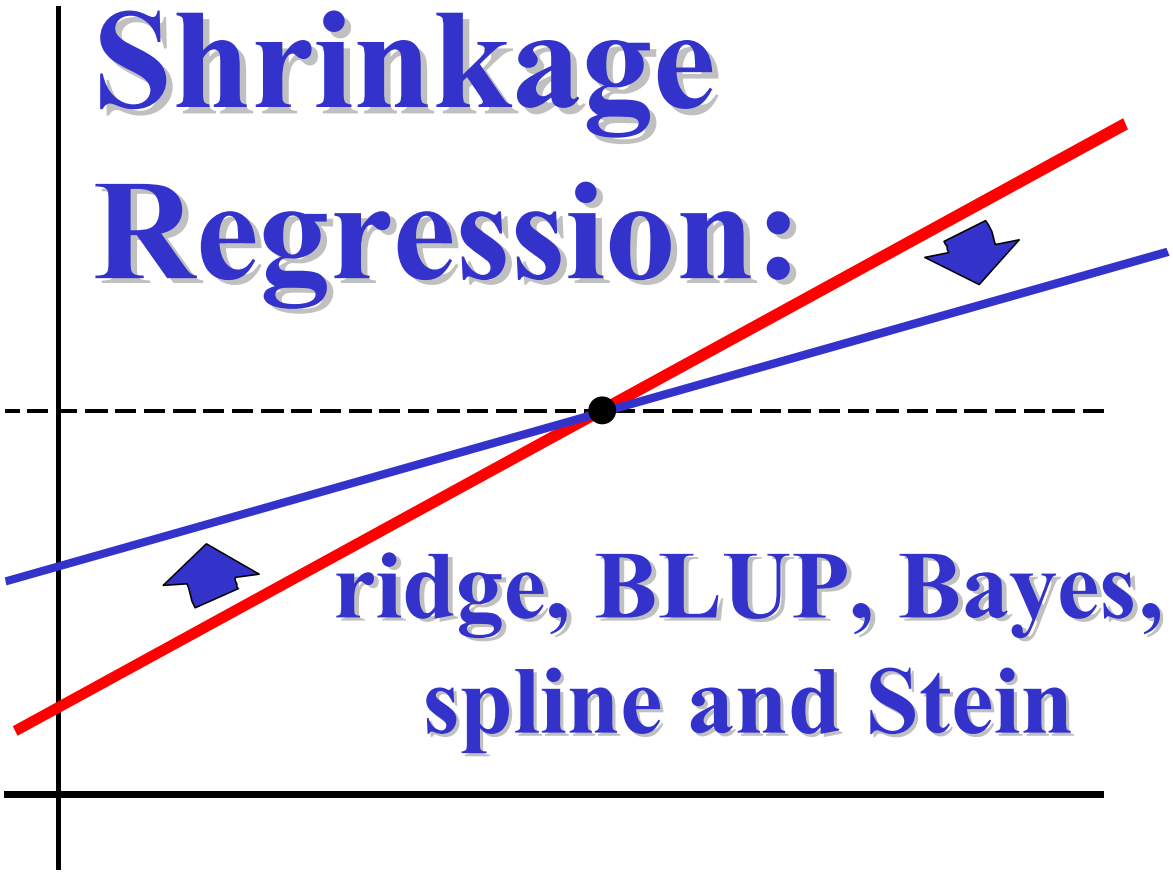
Leamer, E. E. (1978). "Regression selection strategies and revealed priors." **Journal of the American Statistical Association** 73, 580-587.

Obenchain, R. and Vinod, H. (1974). "Estimates of partial derivatives from ridge regression on ill-conditioned data." **NBER-NSF Seminar on Bayesian Inference in Econometrics**, Ann-Arbor, Michigan.

Smith, A. F. M. and Goldstein, M. (1975). "Ridge regression: some comments on a paper of Conniffe and Stone." **The Statistician**, 24, 61-66.

Stone, J. and Conniffe, D. (1973). "A critical view of ridge regression." **The Statistician**, 22, 181-187.

Swamy, P. A. V. B., Rappoport Paul N. (1975). “Relative efficiencies of some simple Bayes estimators of coefficients in dynamic models - I." **Journal of Econometrics** 3, 273-296. 1976.



Chapter 09: Computationally Intense Methods

Bob Obenchain, Ph.D.
softR$_X$ freeware
13212 Griffin Run
Carmel, Indiana 46033-8835

Chapter 9: Computationally Intense Methods (Errors-in-Variables, Resampling and Robustness)

The methodologies discussed in this chapter focus attention upon possible idiosyncrasies in observations of the independent X variables and/or the dependent Y variable of regression models. In addition to observational errors in the Y variable and possible ill-conditioning (high intercorrelations) among X variables that were considered in previous chapters, here we also consider data idiosyncrasies such as imprecise X values as well as "influential" observations (caused by outlying response values and/or high leverage regressor combinations.) Numerical algorithms commonly used to treat these sorts of idiosyncrasies tend to be computationally intense, and any data idiosyncracies detected by these methods can end up being either exploited or simply ignored!

A fitted regression solution is affected not only by (i) the available data but also by one's choices of (ii) statistical model, (iii) estimation methodology, and (iv) computational algorithms. Regression practitioners are well advised to explore the limitations imposed on accuracy and relevance of their fits by uncertainty stemming from all four of these sources. However, especially for the techniques discussed in this chapter, the final refrain from the Ballade of Multiple Regression, Corlett(1963), may well be particularly cogent:

"Your optimum only is bonum
For the data you've fitted it to!"

We start with two sections related to "errors-in-variables" models under which least-squares estimates are biased and inconsistent. Section §9.1 shows how random data un-rounding can suggest shrinkage to improve stability of estimates. But the maximum likelihood methods for multivariate normal "structural" models of section §9.2 suggest, instead, coefficient expansions to "correct-for-attenuation" resulting from uncertainty in predictor variables.

Next, we review iterative methods which, although traditionally applied to regressor variable subsetting or robust fitting, can also be used in shrinkage regression estimation. Section §9.3 discusses methods of cross-validation for choice and assessment of shrinkage. And "robust" methods that down-weight certain types of otherwise "influential" observations are described in section §9.4.

9.1 Regressor Perturbations After the Last Decimal Place

and the Perturbation-Limit

This section discusses the main concepts introduced by Beaton, Rubin and Barone(1976) in their re-analysis of the infamous, ill-conditioned dataset of Longley(1967). Longley had used a six regressor model for U.S. economic results for the sixteen years from 1947 to 1962 to illustrate, rather dramatically, that the least-squares estimates from several software packages (apparently in widespread use at the time of his article) yielded estimates of poor numerical accuracy. As a direct result, being able to produce accurate results on the Longley “benchmark" is now widely considered some sort of minimal litmus-test for commercial statistical packages. After all, simple computational precautions like (internal) centering and scale-standardization of variables, calculating the singular-value-decomposition of X (rather than the eigen-decomposition of X^TX), and/or use of double precision suffice to “pass" this test.

I feel that the revelations of Beaton, Rubin and Barone(1976) were even more dramatic than those of Longley(1967). They showed (i) that small data perturbations (beyond the last decimal place reported in the Longley data) can yield even larger numerical changes in least-squares coefficient estimates than those from the computational algorithms Longley(1967) found to be least accurate. And they also showed (ii) that there is a well defined sense in which some of the estimates found to be least numerically accurate by Longley(1967) are actually more “reasonable" than the most accurate estimates! The following is a summary of the insights provided by Beaton, Rubin and Barone(1976) [B-R-B].

In a given dataset, such as the Longley(1967) benchmark, we may presume that the reported values are absolutely accurate as far as they go. In reality, the reported values have usually been rounded, possibly in some quite subjective way. For example, the Gross National Product (GNP) for 1947 of Longley(1967) was $X_2 = 234{,}289$ (or $234,289,000,000.00.) Rather than being precisely this value, the true GNP for 1947 may well have been almost any value between $234,288,500,000.00 and $234,289,499,999.99. Even a time-trend variable like Longley's $X_6 = \text{YEAR}$ is not a precise value in the sense that different years can contain different numbers of business days.

Now consider adding uniformly-distributed pseudo-random numbers on $[-0.5, +0.5)$, starting in the digit following the last published X digit. Small “errors" like these may seem almost trivial. After all, datasets generated in this way would be identical to the given data when rounded to the given number of digits. In this sense then, the un-rounded data are just as likely to be the exact data as are the given, rounded data.

Six of the 15 pairwise correlations among the six regressor variables in Longley's benchmark exceed $+0.98$. Because two of Longley's six X variables are reported with six significant figures and all have at least 3 significant figures, adding uniformily-distributed random deviates confined to the digits following the last reported decimal place assures that each set of perturbed regressor coordinates will remain intensely ill-conditioned and, thus, numerically unstable. And B-R-B found that the numerically accurate solution for the rounded (unperturbed) Longley data is “nowhere near the center of the distribution of a large number of presumably equally plausible [perturbed] solutions."

Consider the regression model $y = T \cdot \beta + \epsilon$ where (i) y is the $N \times 1$ vector of observations on the dependent variable, (ii) T is the $N \times p$ matix of true regressor values, (iii) β is the $p \times 1$ vector of parameters to be estimated, and (iv) ϵ is a $N \times 1$ "catchall" vector for the unpredictable portion of y. The least-squares estimate of β would then be $\widehat{\beta} = (T^T T)^{-1} T^T y$. Alternatively, letting X denote the observed, rounded version of T and writing $E = T - X$ for the difference between T and X, we have

$$\widehat{\beta} = [(X + E)^T (X + E)]^{-1} (X + E)^T y = [X^T X + E^T X + X^T E + E^T E]^{-1} (X^T y + E^T y).$$

Unfortunately, the numerical values of the T and E matrices are unknown. However, we may simulate the true least-squares estimates using a sequence of E matrices that represent statistically independent, uniformly-distributed data un-roundings. It then follows, for example, that $E(E) = 0$, $E(E^T X) = E(X^T E) = 0$ for fixed X, and $E(E^T y) = 0$ for fixed y. From similar expressions for fixed moments of order four or less, it follows that

$$E(\widehat{\beta}) = [X^T X/N + E(E^T E/N)]^{-1} X^T y/N, \qquad \{ 9.1 \}$$

where the expected value of $E^T E/N$ is the diagonal matrix of known variances of the regressor variable un-roundings. (Because the uniform distribution on $[-0.5, +0.5)$ has mean zero and variance $1/12 = 0.083\overline{3}$, un-rounding added following the d-th place after the decimal would have variance $0.083\overline{3} \times 10^{-2d}$.)

B-R-B call { 9.1 } the P-lim (perturbation limit) of their simulations and note that it is a Hoerl-Kennard(1970) ridge (shrinkage) estimator like that of our equation { 3.6 }, except that the diagonal elements of the $E(E^T E/N)$ matrix may not be all equal. Because $X^T X/N$ will have large off-diagonal elements (and at least some numerically small eigenvalues) when regressors are highly intercorrelated, the addition through $E(E^T E/N)$ of even relatively small (but positive) numerical values to the diagonal of $X^T X/N$ can create dramatic numerical changes in $E(\widehat{\beta})$ relative to the un-perturbed least-squares vector, $[X^T X/N]^{-1} X^T y/N$. In particular, these numerical changes can include even (i) changes in the numerical signs of some stabalized coefficients and (ii) changes of more than 5 digits in the second most significant figure of many coefficients.

In summary, then, Beaton, Rubin and Barone(1976) argue that the numerically most accurate least-squares solution corresponding to an ill-conditioned dataset can be extreme and implausible relative to the general distribution of solutions that can result from random data un-rounding. Furthermore, the (P-lim) "center" of this un-rounding distribution for coefficeints corresponds to a shrunken version of the un-perturbed least-squares coefficient estimates.

9.2 Multivariate Normal Errors-in-Variables Analyses

The presence of unknown measurement errors in the observed values of regressor variables leads to distortions in regression coefficient estimates; see, for example, Cochran(1968), Fuller(1987) and Gleser(1992). In the most simple case of a single (nonconstant) regressor

variable, $P = 1$, it is well known that the expected value of the resulting slope estimate is reduced in absolute value relative to the slope expected from regression on x-values free of measurement error; this effect is commonly called attenuation, as in Fuller and Hidiroglou(1978). Here, we outline some of the most obvious effects of measurement errors on coefficient estimates of multiple regression models and show that the resulting bias frequently corresponds to various forms of shrinkage of expected coefficient values. Correcting for this sort of bias in least-squares estimates motivates certain expansions of coefficient estimates, with corresponding inflations in their estimated variances.

Linear Errors-In-Variables Regression, EIVR, models (before centering) are of the form:

$$y = \tau + \epsilon \qquad \{ 9.2 \}$$

where

y is the $N \times 1$ vector of observations on the dependent variable,
ϵ is the $N \times 1$ vector of errors in the dependent variable,

and

$\tau = 1 \cdot \alpha + T \cdot \beta$ for $X = T + F$.

9.3 Cross-Validation, Bootstrapping, and Predictive Sample Reuse

9.3.1 Allen's PRESS Criterion.

9.3.2 A Rotation-Invariant Prediction Criterion.

9.4 Iterative Re-Weighting Methods

The results developed in Section §2.11, Analyses of Residuals, are sufficiently detailed to make them immediately applicable to most iterative robust-regression algorithms.

References for Chapter Nine

Allen, D. M. (1974). "The relation between variable selection and data augmentation and a method for prediction." **Technometrics** 16, 125-127.

Andrews, D. F. (1974). "A robust method for multiple linear regression." **Technometrics** 16, 523-531.

Askin, R. G. and Montgomery, D. C. (1980). "Augmented robust estimators." **Technometrics** 22, 333-341.

Atkinson, A. C. (1986). "Masking unmasked." **Biometrika** 73, 533-541.

Bassett, G. and Koenker, R. (1978). "Asymptotic theory of least absolute error regression." **Journal of the American Statistical Association** 73, 618-622.

Beaton, A. E. and Tukey, J. W. (1974). "The fitting of power series, meaning polynomials, illustrated using band-spectroscopic data." **Technometrics** 16, 147-179.

Beaton, A. E., Rubin, D. B. and Barone, J. L. (1976). "The acceptability of regression solutions: another look at computational accuracy." **Journal of the American Statistical Association** 71, 158-168.

Berk, K. N. (1978). "Comparing subset selection procedures." **Technometrics** 20, 1-6.

Chambers, J. M. (1972). "Stabilizing linear regression against observational error in the independent variates." Bell Telephone Laboratories, Murray Hill, NJ.

Chambers, J. M. (1973). "Linear regression computations: some numerical and statistical aspects." **Proceedings of the 39th Session of the International Statistical Institute**, 45, 245-254.

Cochran, W. G. (1968). "Errors of measurement in statistics." **Technometrics** 10, 637-666.

Cook, R. D. (1977). "Detection of influential observations in regression." **Technometrics** 19, 15-18.

Cook, R. D. and Weisberg, S. (1989). "Regression diagnostics with dynamic graphics." **Technometrics** 31, 277-291. [discussion 293-311.]

Corlett, T. (1963). "Ballade of multiple regression." **Applied Statistics** 12, 145.

Dallal, G. E., Rousseeuw, P. J., Leroy, A. M. and van Zomeren, B. C. (1991). **LMS: Least Median of Squares Regression**. FORTRAN Software for IBM-compatible Personal Computers.

Denby, L. and Mallows, C. L. (1977). "Two diagnostic displays for robust regression analysis." **Technometrics** 19, 1-13.

Fuller, W. A. and Hidiroglou, M. A. (1978). "Regression estimation after correcting for attenuation." **Journal of the American Statistical Association**, 73, 99-104.

Fuller, W. A. (1987). **Measurement Error Models**. New York, NY: John Wiley.

Gleser, L. J. (1992). "The importance of assessing measurement reliability in multivariate regression." **Journal of the American Statistical Association** 87, 696-707.

Gleser, L. J., Carroll, R. J. and Gallo, P. P. (1987). "The limiting distribution of least squares in an errors-in-variables model." **The Annals of Statistics** 15, 220-233.

Hawkins, D. M., Bradu, D. and Kass, G. V. (1984). "Location of several outliers in multiple-regression data using elemental sets." **Technometrics** 26, 197-208.

Henderson, H. V. and Velleman, P. (1981). "Building multiple regression models interactively." **Biometrics** 37, 391-411.

Hettmansperger, T. P. and McKean, J. W. (1977). "A robust alternative based on ranks to least squares in analyzing general linear models." **Technometrics** 19, 275-284.

Hettmansperger, T. P. and McKean, J. W. (1978). "Statistical inference based on ranks." **Psychometrika** 43, 69-79.

Hocking, R. R. (1972). "Criteria for selection of a subset regression: which one should be used?" **Technometrics**, 14, 967-970.

Hocking, R. R. (1976). "The analysis and selection of variables in linear regression." **Biometrics** 32, 1-49.

Holland, P. (1973). "Weighted ridge regression: combining ridge and robust regression methods." Working Paper #11, National Bureau of Economic Research, Cambridge, Massachusetts.

Kapenga, J. A. and McKean, J. W. (1988). "The vectorization of algorithms for R-estimates in linear models." **Proceedings of the 19th Symposium on the Interface: Computer Science and Statistics**, R. M. Heiberger, ed. 502-506.

Kapenga, J. A., McKean, J. W. and Vidmar, T. J. (1988). **RGLM: A Robust General Linear Model Package**, Version ASA-1.01. Kalamazoo, Michigan: Western Michigan University and The Upjohn Company.

Krasker, W. S. and Welsch, R. E. (1982). "Efficient bounded influence regression estimation." **Journal of the American Statistical Association** 77, 595-604.

Mason, R. L. and Gunst, R. F. (1985). "Outlier-induced collinearities." **Technometrics** 27, 401-407.

McKean, J. W. and Hettmansperger, T. P. (1976). "Tests of hypotheses of the general linear model based on ranks." **Communications in Statistics** A5, 693-709.

McKean, J. W. and Hettmansperger, T. P. (1980). "A robust analysis of the general linear model based on one-step R-estimates." **Biometrika** 65, 571-579.

McKean, J. W., Sheather, S. J. and Hettmansperger, T. P. (1976). "Regression diagnostics for rank-based methods." **Journal of the American Statistical Association** 85, 1018-1028.

Pariente, S. and Welsch, R. E. (1977). "Ridge and robust regression using parametric linear programming." Working Paper. Alfred P. Sloan School of Managemant, MIT, Cambridge, Massachusetts.

Pichard, R. R. and Berk, K. N. (1980). "Data splitting." **The American Statistician** 44, 140-147.

Pichard, R. R. and Cook, R. D. (1984). "Cross-validation of regression models." **Journal of the American Statistical Association** 79, 575-583.

Ramsay, J. O. (1977). "A comparative study of several robust estimates of slope, intercept, and scale in linear regression." **Journal of the American Statistical Association** 72, 608-615.

Roecker, E. B. (1991). "Prediction error and its estimation for subset-selected models." **Technometrics** 33, 459-468.

Rousseeuw, P. J. (1984). "Least median of squares regression." **Journal of the American Statistical Association** 79, 871-880.

Rousseeuw, P. J. and van Zomeren, B. C. (1990). "Unmasking multivariate outliers and leverage points." **Journal of the American Statistical Association** 85, 633-639. [discussion 640-651.]

Walker, E. (1989). "Detection of collinearity-influential observations." **Communications in Statistics** A18, 1675-1690.

Wu, C. F. J. (1986). "Jackknife, bootstrap and other resampling plans in regression analysis." **The Annals of Statistics** 14, 1261-1295.

Additional Reading for Chapter Nine

Afifi, A. A. and Azen, S. P. (1972). **Statistical Analysis: A Computer Oriented Approach.** New York: Academic Press.

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). **Regression Diagnostics**. New York: John Wiley and Sons.

Chambers, J. M., Cleveland, W., Kleiner, B. and Tukey, P. (1983). **Graphical methods for data analysis**. Monterey, California: Wadsworth and Brooks-Cole.

Cook, R. D. and Weisberg, S. (1982). **Residuals and Influence in Regression**. New York: Chapman and Hall.

Daniel, C. and Wood, F. S. (1971). **Fitting Equations to Data: Computer Analysis of Multifactor Data for Scientists and Engineers**. New York: Wiley-Interscience.

Draper, N. R. and Smith, H. (1981). **Applied Regression Analysis**, Second Edition. New York: John Wiley and Sons.

Hampel, F., Ronchetti, E., Rousseeuw, P. and Stahel, W. (1986). **Robust Statistics**. New York: John Wiley and Sons.

Huber, P. J. (1981). **Robust Statistics**. New York: John Wiley and Sons.

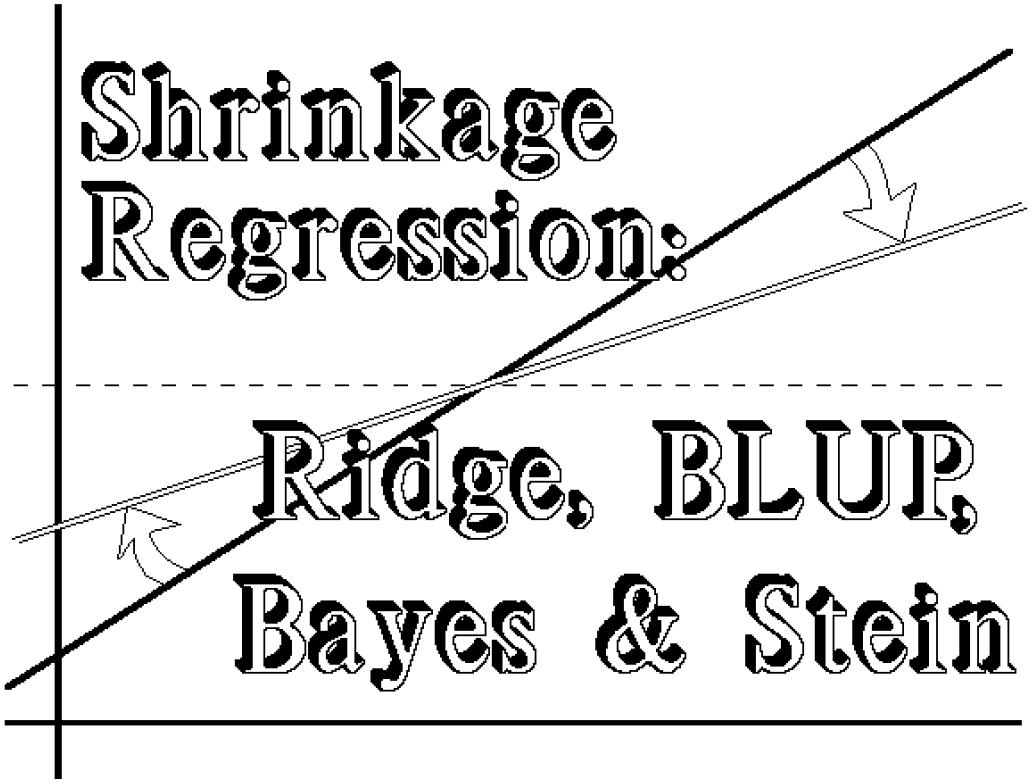
L'Ecuyer, P. (1988). "Efficient and portable combined random number generators." **Communications of the ACM** 31, 742-749,774.

Park, S. K. and Miller, K. W. (1988). "Random number generators: good ones are hard to find." **Communications of the ACM** 31, 1192-1201.

Press, W. H., Flannery, B. P., Teukolsky, S.A., and Vetterling, W. T. (1988). **Numerical Recipes in C: The Art of Scientific Computing**. [especially Chapter 7: Random Number Generation.] Massachusetts: Cambridge University Press. {Code Copyright 1985, 1987 by Numerical Recipes Software P. O. Box 243, Cambridge, MA 02238.}

Rousseeuw, P. J. and LeRoy, A. M. (1987). **Robust Regression and Outlier Detection**. New York: John Wiley and Sons.

Wichman, B. A. and Hill, I. D. (1982). "Algorithm AS 183: An efficient and portable pseudo-random number generator." **Applied Statistics** 31, 188-190.



Chapter 10: Historical Topics

Bob Obenchain, Ph.D.
softR$_X$ freeware
5261 Woodfield Drive North
Carmel, Indiana 46033-8795

Chapter 10: Topics of Historical Interest, Heuristic Arguments & Common Misconceptions.

10.1 The Contributions of Hoerl and Kennard.

10.1.1 Characterization of the “Ordinary" Ridge Shrinkage Path.

10.1.2 The Hoerl-Kennard “too-longness" argument.

10.1.3 The Hoerl-Kennard “existance theorem."

10.1.4 Heuristic “Fixed-Point" Arguments.

10.2 The distinction between “weak" data and “ugly" data.

10.3 The “Chain-Rule" Argument of Obenchain-Vinod.

My joint paper with Rick Vinod [Obenchain and Vinod(1974)] was my first work devoted exclusively to shrinkage regression. Unfortunately, this paper turned out to be highly “controversial" in several senses. For example, we had to revise it twice just to get it released for publication from Bell Telephone Laboratories, then we performed another major revision between submissions to two different journals. But it was never accepted for publication!

As is clear from the style and grammar of these early manuscripts, I (rather than Rick) did most of the writing, but our original intent was simply to explain/defend a concept due to Rick [Vinod(1971)] called the **Chain-Rule Argument**. Much of the notation and terminology of these manuscripts was new/innovative at that time; many of the fundamental relationships described in Chapters 2 and 3 of this book first appeared in these manuscripts. Furthermore, the concepts of PCSA and of plotting TRACES versus MCAL were first introduced there. Ultimately, just about everything from these papers, **except** the chain-rule argument, got published in Obenchain(1975,1976) and/or in Vinod(1976).

shrunken coefficients as constrained partial derivatives.

Cox(1971) argues that a reparameterization of Scheffe coefficients for mixture experiments is desirable because it allows the new coefficients to be interpreted as directional derivatives. Of course, Cox is almost surely assuming that the data to be fitted come from a “designed experiment."

10.4 “Fictitious Data Augmentation" Arguments.

10.5 Preliminary-Test Estimation.

Furthermore, my references to “preliminary testing" came, politically speaking, at a most inappropriate time. The highly influential work of Bock, Yancey, and Judge(1973) was being widely interpreted, around then, as sounding a “death nell" for preliminary test methodology!

In my paper, I had adopted what a 1950's television critic described as the “BEVERLY HILLBILLY'S STRATEGY," which was to... “Aim Low, and Hit Your Mark." I had devised a statistically valid testing procedure for what McDonald suggested I call the UNIFORM SHRINKAGE HYPOTHESIS, under which least squares regression coefficients have mean-squared-error optimal relative magnitudes. This test (and its associated approximate F-distribution) provided ample power to detect ill-conditioned regression problems and, thereby, “reject" least-squares coefficients on the grounds that their relative magnitudes might be potentailly misleading. Apparently, my testing procedure also provides almost unlimited potential for ABUSE.

10.6 Latent-Root Regression Methods.

10.7 Adjusting Correlations Among Regression Coefficients.

Of all the papers I ever “volunteered" to review for internal clearance by Bell Telephone Laboratories, Tukey(1975) provides my most vivid memories of the strengths and weaknesses of a “peer-review" process. I suspect that John Tukey was least happy when I described the methodology of his Section 6 as a “new form of ridge regression."

10.8 Best Linear and Quadratic Estimators of One.

References for Chapter Ten

Allen, D. M. (1974). "The relation between variable selection and data augmentation and a method for prediction." **Technometrics** 16, 125-127.

Bacon, R. W. and Hausman, J. A. (1974). "The relationship between ridge regression and the minimum mean square error estimator of Chipman." **Oxford Bulletin of Economics and Statistics**, 36, 115-124.

Berk, K. N. (1978). "Comparing subset selection procedures." **Technometrics** 20, 1-6.

Bock, M., Yancey, T. A. and Judge, G. G. (1973). "The statistical consequences of preliminary test estimators in regression." **Journal of the American Statistical Association** 68, 109-116.

Box, G. E. P. (1966). "Use and abuse of regression." **Technometrics**, 8, 625-629.

Cox, D. R. (1971). "A note on polynomial response functions for mixtures." **Biometrika** 58, 155-159.

Crone, L. (1972). "The singular value decomposition of matrices and cheap numerical filtering of systems of linear equations." **Journal of the Franklin Institute** 294, 133-136.

Dempster, A. P. (1973). "Alternatives to least squares in multiple regression." **Multivariate Statistical Inference**. Eds. Kabe, D. G. and Gupta, R. P. Amsterdam: North-Holland Publishing Company, pp25-40.

Draper, N. R. and Van Nostrand, R. C. (1977a). "Ridge regression and James-Stein estimation: review and comments." **Technometrics** 21, 451-466.

Egerton and Laycock (1981). **Technometrics** 23, 155-158.

Farebrother, R. W. (1975). "The minimum mean square error linear estimator and ridge regression." **Technometrics** 17, 127-128.

Harville, D. A. (1986). "Using least squares software to compute combined intra-interblock estimates of treatment contrasts," **The American Statistician**, 40, 153-157.

Hawkins, D. M. (1975). "Relations between ridge regression and eigen-analysis of the augmented correlation matrix." **Technometrics**, 17, 477-480.

Hemmerle, W. J. (1975). "An explicit solution for generalized ridge regression." **Technometrics**, 17, 309-314.

Hill, M. A. (1975). "Ridge regression using BMDP2R." **BMD Communications** No. 3 (UCLA Health Sciences Computing Facility) Feb. 75, 1-2.

Hocking, R. R. (1972). "Criteria for selection of a subset regression: which one should be used?" **Technometrics**, 14, 967-970.

Hocking, R. R. (1976). "The analysis and selection of variables in linear regression." **Biometrics** 32, 1-49.

Hoerl, A. E. (1962). "Application of ridge analysis to regression problems." **Chemical Engineering Progress** 58, 54-59.

Hoerl, A. E. and Kennard, R. W. (1970a). "Ridge regression: biased estimation for nonorthogonal problems." **Technometrics** 12, 55-67.

Hoerl, A. E. and Kennard, R. W. (1970b). "Ridge regression: applications to nonorthogonal problems." **Technometrics** 12, 69-82.

Hoerl, A. E., Kennard, R. W. and Baldwin, K. F. (1975). "Ridge regression: some simulations." **Communications in Statistics** A4, 105-123.

Hoerl, A. E. and Kennard, R. W. (1975). "A note on a power generalization of ridge regression." **Technometrics**, 17, 269.

Hoerl, A. E. and Kennard, R. W. (1976). "Ridge regression: iterative estimation of the biasing parameter." **Communications in Statistics** A5, 77-88.

Lowerre, J. M. (1974). "On the mean square error of parameter estimates for some biased estimators." **Technometrics**, 16, 461-464.

Marquardt, D. W. and Snee, R. D. (1975). "Ridge regression in practice." **The American Statistician**, 29, 3-19.

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

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

McDonald, G. C. and Schwing, R. C. (1973). "Instabilities of regression estimates relating air pollution to mortality." **Technometrics**, 15, 463-481.

Obenchain, R. L. and Vinod, H. D. (1974). "Estimates of partial derivatives from ridge regression on ill-conditioned data." Presented at the **NBER-NSF Seminar on Bayesian Inference in Econometrics**. Ann-Arbor, Michigan.

Piegorsch, W. W. and Casella, G. (1989). "The early use of matrix diagonal increments in statistical problems." **Siam Review** 31, 428-434.

Pichard, R. R. and Berk, K. N. (1980). "Data splitting." **The American Statistician** 44, 140-147.

Pichard, R. R. and Cook, R. D. (1984). "Cross-validation of regression models." **Journal of the American Statistical Association** 79, 575-583.

Roecker, E. B. (1991). "Prediction error and its estimation for subset-selected models." **Technometrics** 33, 459-468.

Strawderman, W. E. (1978). "Minimax adaptive generalized ridge regression estimators." **Journal American Statistical Association** 73, 623-627.

Theil, H. (1971). **Principles of Econometrics**. Amsterdam: North Holland Publishing Company.

Tukey, J. W. (1975). "Instead of Gauss-Markov Least Squares; What?" **Applied Statistics**, ed. R. P. Gupta. New York: North Holland Publishing Company.

Vinod, H. D. (1971). "Regression with sign restricted coefficients." Presentation at the Annual Meeting of the Econometric Society, New Orleans.

Vinod, H. D. (1972). "Nonhomogeneous production functions and applications to telecommunications." **Bell Journal of Economics and Management Science** 3, 531-543.

Vinod, H. D. (1974). "Ridge estimation of a trans-log production function." **Business and Economic Statistics Section, Proceedings of the American Statistical Association**.

Vinod, H. D. (1976a). "Applications of new ridge regression methods to a study of Bell System scale economies." **Journal of the American Statistical Association** 71, 835-841.

Vinod, H. D. (1976b) "Canonical ridge and econometrics of joint production." **Journal of Econometrics** 4, 147-166.

Vinod, H. D. (1976c). "Simulation and extension of a minimum mean squared estimator in comparison with Stein's." **Technometrics** 18, 491-496.

Vinod, H. D. (1978). "A survey of ridge regression and related techniques for improvements over ordinary least squares." **The Review of Economics and Statistics** 60, 121-131.

Vinod, H. D. and Ullah, A. (1981). **Recent Advances in Regression Methods** New York: Marcel Dekker.

Webster, J. T., Gunst, R. F. and Mason, R, L. (1974). “Latent root regression analysis." **Technometrics**, 16, 513-522.

Wu, C. F. J. (1986). “Jackknife, bootstrap and other resampling plans in regression analysis." **The Annals of Statistics** 14, 1261-1295.