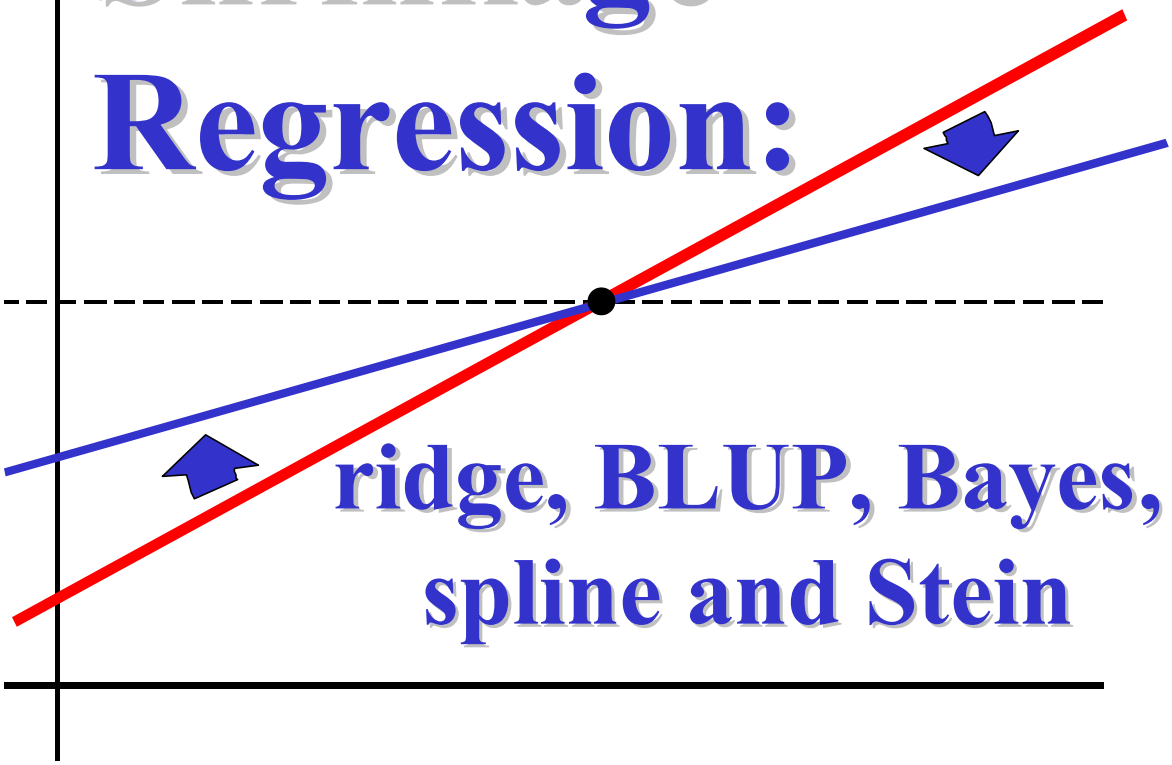
Shrinkage Regression:



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.

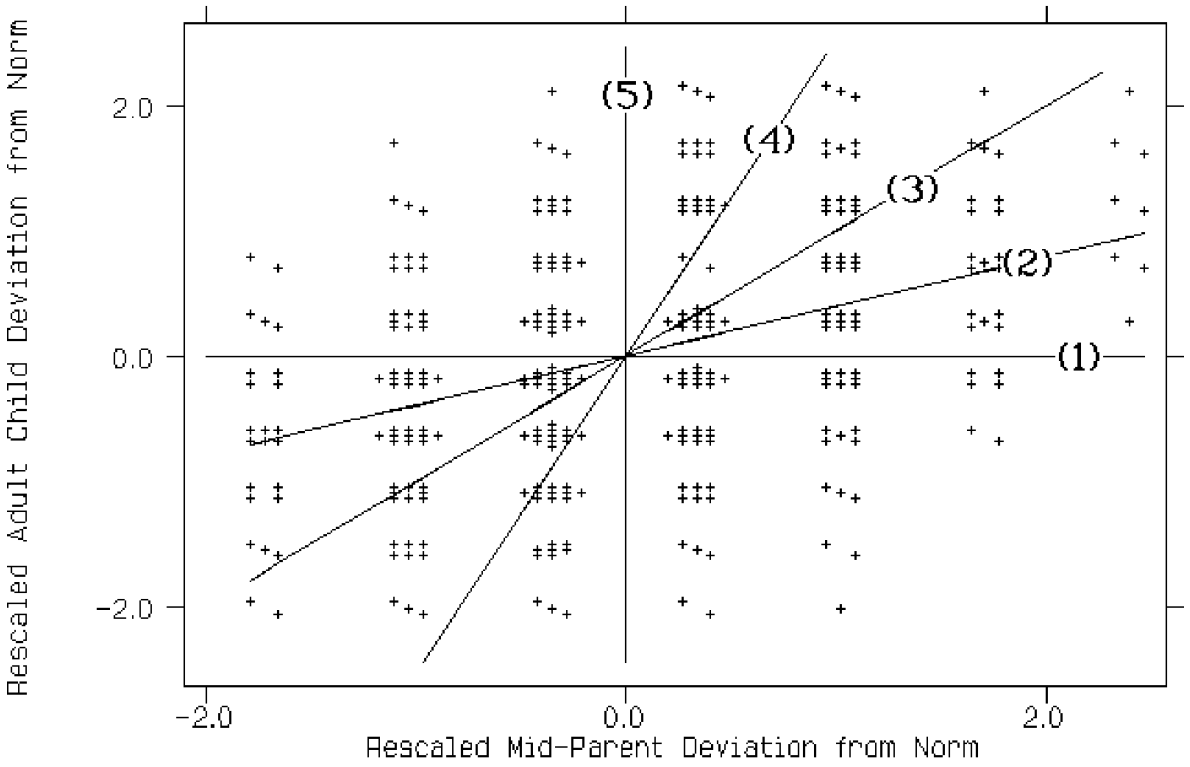
1.1 Galton's "Shrinkage" Interpretation of Regression

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 ...

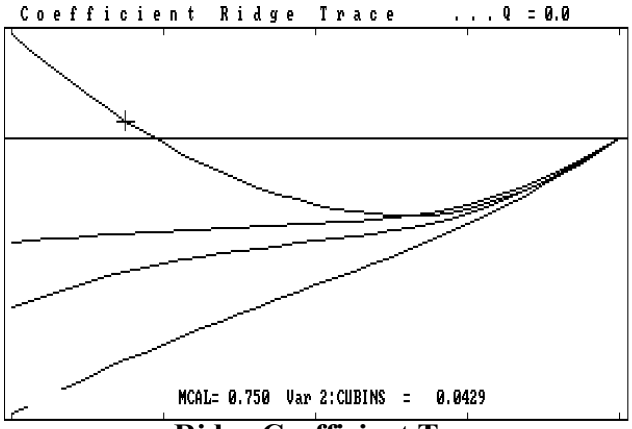
Table 1.3: Shrinkage Shape Summary

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	

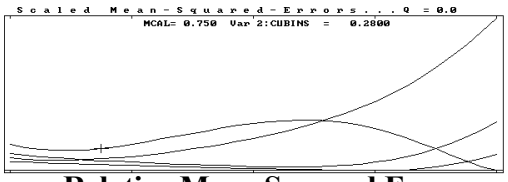
This analysis favors a shrinkage extent of about MCAL $= 2$ along the path of shape QPAR $= -1$. And we have already ruled out the uniform (QPAR $= +1$) shape as being of no interest in this application. Suppose we decide to actually explore the path of shape QPAR $= 0$ because (i) this corresponds to 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...

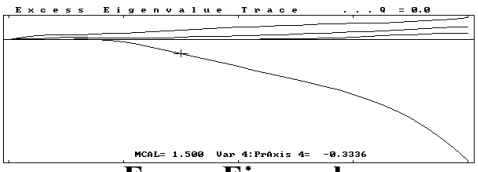
Figure 1.2: Shrinkage TRACE Displays



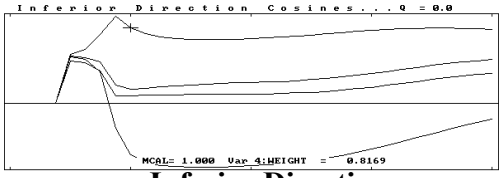
Ridge Coefficient Trace



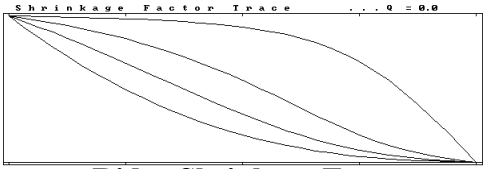
Relative Mean Squared Errors



Excess Eigenvalues



Inferior Direction

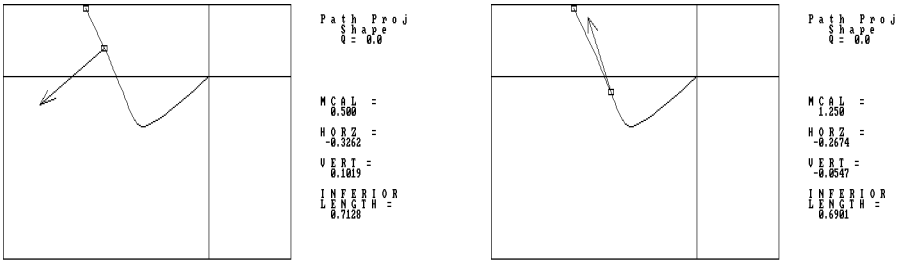


Ridge Shrinkage Factors

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

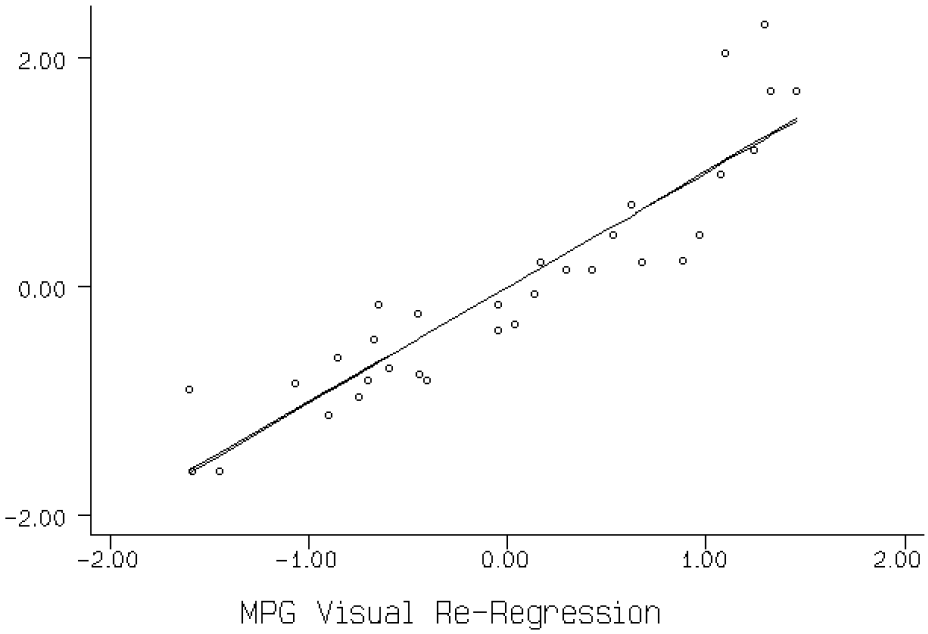
MCAL	CLIK	EBAYS	RCOEF
0.000	+infinity	+infinity	+infinity
0.125	90.9	17.3	17.4
0.250	29.6 <(2/R)	15	15.2
0.375	15.1	13.9	14.1
0.500	10.7	13.3	13.6
0.625	9.95 <<<	12.9	13.3
0.750	10.7	12.8 <<<	13.2 <<<
0.875	12.1	12.8	13.2
1.000	13.9	13	13.4
2.000	30.6	19.6	18.6
3.000	45.8	52.5	35.4
4.000	60.4	151	60.4

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.