ridge (shrinkage) regression in Stata

by

Bob Obenchain, Ph.D.

softRX freeware / Risk Benefit Statistics LLC

1 Introduction

Ridge regression is a graphically oriented methodology for analysis of ill-conditioned (multicollinear) regression models. Ridge methods tend to be computationally intensive, especially when normal-theory maximum likelihood estimation techniques are incorporated to provide objective information about the most appropriate form and extent of shrinkage. This paper presents an overview of ridge concepts along with five Stata programs to monitor the effects of shrinkage.

Key Words and Phrases: classical, fixed coefficients; empirical Bayes; random coefficients; 2-parameter ridge family; multicollinearity allowance axis; true risks; simulated losses.

1.1 Ill-Conditioning and Ridge Regression

Fitting of models to ill-conditioned data collected retrospectively poses serious obstacles to multiple regression practitioners, particularly in such fields as economics where interest can focus on the relative sizes of estimated coefficients. Consider the classical multiple regression model

$$y = 1\mu + X\beta + \epsilon \tag{1}$$

where y is a $n \times 1$ vector of observations on the response variable, μ is the unknown intercept, X is a $n \times p$ matrix containing coordinates for $p \geq 2$ non-constant predictor variables, β is a $p \times 1$ vector of unknown coefficients, and ϵ is a $n \times 1$ vector of unobserved, normally-distributed disturbance terms

$$\epsilon \sim N(0, \sigma^2 I) \ . \tag{2}$$

If the predictor variables are "centered" by subtracting off their observed means and the resulting X matrix of explanatory variables is of full column rank, then the maximum likelihood estimate of β is the least-squares solution

$$\widehat{\beta} = (X'X)^{-1}X'y \ . \tag{3}$$

It is then straightforward to show that

$$\widehat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1}) \ . \tag{4}$$

Problems arise when X is ill-conditioned. Numerical ill-conditioning occurs when exact linear relationships exist between, say, the $i-$th and $j-$th explanatory variables. That is,

$$x_i = a + bx_j \ , \tag{5}$$

where a and b are constants. In this case, $X'X$ is singular, and $\hat{\beta}$ is not uniquely determined.

More commonly, two or more X variables are highly correlated, and $X'X$ approaches singularity. In this situation, $\widehat{\beta}$ is unique but is imprecisely estimated. In other words, the relative magnitudes of the elements of $\widehat{\beta}$ may be distorted (they may even have "wrong" numerical signs) because the fitted coefficients are also highly correlated. As a result of this "statistical" ill-conditioning, elements of $\hat{\beta}$ or certain linear combinations may be insignificant primarily because their variances are relatively large.

The topic of ill-conditioned regression models is one of the most thoroughly researched problems in statistics, and *ridge regression* is one approach that has been proposed to treat the symptoms of ill-conditioning. Ridge estimators shrink the estimated coefficient vector, $\hat{\beta}$, and thus provide biased estimates of β. But variance is also reduced by shrinking, so ridge estimators can achieve lower Mean Squared Error (MSE) risk than least-squares.

An intuitive way to treat ill-conditioning is to increase the diagonal elements of

$X'X$ before attempting to invert this inner products matrix and form $\hat{\beta}$ via equation (3), Piegorsch and Casella(1989). Anyway, the original ridge estimator of Hoerl(1962) was

$$\beta^* = (X'X + kI)^{-1}X'y \tag{6}$$

where k is a small, positive constant.

Interest in ridge regression was sparked by Hoerl and Kennard(1970a,b) when they suggested plotting the p elements of β^* as a function of k in a graphical display called the *ridge trace*. They observed that the relative magnitudes of the elements of β^* tend to "stabilize" as k increases and over-optimistically conjectured that is "easy" to pick an extent of shrinkage yielding lower MSE than least squares. For more than twenty years now, a storm of criticisms, alternative proposals for choice of k, and ridge simulation studies have appeared in statistical literature. If any sort of consensus has emerged, it may well be that (a) ridge methods tend to shrink much too much to be anywhere close to being minimax rules [i.e. you can end up either winning big by reducing MSE or else loosing big by increasing MSE] and (b) the "generalized cross validation" method of Golub, Heath and Wahba(1979) for picking an appropriate extent of shrinkage is a consistent high-performer in simulation studies.

Classical, normal-theory maximum likelihood estimation in generalized ridge regression has been a research interest of mine since 1973. In my first published ridge

paper, Obenchain (1975), I derived general equations for "likelihood monitoring" that generated little interest, apparently due to their complexity. However, Gibbons(1981) did evaluate this "O-method" and found that it out-performed Golub, Heath and Wahba(1979) "generalized cross validation" in her favorable-case MSE simulations. In Obenchain(1981), I restricted interest to a specific 2-parameter family of generalized ridge estimators, equation (12) below, and derived a closed form expression for the extent of shrinkage along a given ridge path that is most likely to achieve minimum MSE risk; see equations (15) and (16) below. This maximum likelihood approach to shrinkage is fairly conservative in the sense that it reduces the MSE risk by only about 50% even when its $\delta^{MSE} = 0$ [equation(13), below]; but this conservatism also means that the maximum likelihood approach can increase MSE by at most 25% in the least favorable cases, usually somewhere in the $\delta^{MSE} = 0.8$ to $\delta^{MSE} = 0.9$ range. Ridge methods that shrink more aggressively than maximum likelihood tend to either do a little better or else much, much worse on MSE, depending upon whether the application is either favorable or unfavorable to shrinkage, respectively.

Other ridge research efforts of mine, Obenchain(1978,1984), lead to greater understanding of a variety of *multivariate risk* (matrix valued MSE) characteristics of shrinkage estimators, along with corresponding normal-theory maximum likelihood estimates. Like ridge coefficients, these risk estimates can also be plotted in *traces* to display the effects of shrinkage and to help ridge practitioners decide whether to

start shrinking in the first place and, once they start shrinking, where to stop!

1.2 Principal Components and Generalized Ridge Regression

This subsection contains technical details of generalized ridge estimation that may be skipped over on first reading. Here we show (i) how to decompose least squares estimates into uncorrelated components and principal correlations, (ii) that regression on principal components is a special case of generalized ridge regression, and (iii) how ridge estimators shrink least-squares coefficients along the principal axes of the given X coordinates.

Even in cases where X is numerically ill-conditioned, $rank(X) = r < min(p, n-1)$, the singular value decomposition of X can be written as $X = H\Lambda^{1/2}G'$. In this decomposition, H is a $n \times r$ semi-orthogonal matrix of standardized *principal coordinates*, G is a $p \times r$ semi-orthogonal matrix of *principal axis direction-cosines*, and $\Lambda^{1/2}$ is a $r \times r$ diagonal matrix of ordered *singular values*, $\lambda_1^{1/2} \geq \cdots \geq \lambda_r^{1/2} > 0$.

Although the least-squares solution is not uniquely determined when $r < p$, the shortest least-squares coefficient vector is $\hat{\beta} = G\Lambda^{-1/2}H'y \equiv Gc$, where c is the $r \times 1$ vector of *uncorrelated components* of $\hat{\beta}$. Note that

$$c \sim N(\gamma, \sigma^2\Lambda^{-1}) \tag{7}$$

where $\gamma \equiv G'\beta$ are the r unknown *t*rue components of β. The structure of these

uncorrelated components, $c = \Lambda^{-1/2}H'y$, provides key insights into the nature of statistical ill-conditioning:

$$c_i = r_i^o \cdot \sqrt{\frac{y'y}{\lambda_i}} \tag{8}$$

where r_i^o is the *principal correlation* between y and the i-th column of H, the familiar R-squared statistic is $R^2 = r_1^{o2} + \cdots + r_r^{o2}$, and the t-statistic for testing $\gamma_i = 0$ is

$$t_i = \frac{c_i}{\hat{\sigma} \cdot \lambda_i^{-1/2}} = r_i^o \cdot \sqrt{\frac{n-r-1}{(1-R^2)}} \ . \tag{9}$$

Thus the i-th principal correlation, r_i^o, determines whether the i-th component is statistically significant, and yet c_i can be large numerically simply because its λ_i is relatively small rather than because its r_i^o is relatively large!

Linear generalized ridge estimates are of the form

$$\beta^* = G\Delta c = \sum g_i \delta_i c_i \ , \tag{10}$$

where Δ is a $r \times r$ diagonal matrix of *non-stochastic shrinkage factors*, $\delta_1, \cdots, \delta_r$,and g_i is the i-th column of G. Each shrinkage factor lies in the closed interval from zero to one, $0 \leq \delta_i \leq 1$, and the total extent of shrinkage is measured by

$$m = r - \delta_1 - \cdots - \delta_r = rank(X) - trace(\Delta) \ . \tag{11}$$

This m is called the *multicollinearity allowance* ridge parameter, introduced and

discussed in Obenchain and Vinod (1974), Vinod (1976) and Obenchain (1981). Ridge coincides with least squares at $m = 0$ [$\delta_1 = \cdots = \delta_r = 1$], and all ridge coefficients approach zero as m approaches its upper limit of $m = p$ [$\delta_1 = \cdots = \delta_r = 0$].

Regression on principal components is the special case of equation (10) in which each δ_i is either 0 or 1. Standard methods for deciding which δ_i to set equal to zero are: (a) the components with the smallest singular values, $\lambda_i^{1/2}$, or (b) the components with the smallest absolute principal correlations, $|r_i^o|$.

Our primary focus will be on the *2-parameter ridge family* in which the shrinkage factor applied to the i-th uncorrelated component of the least-squares solution is of the general form

$$\delta_i = \lambda_i/(\lambda_i + k\lambda_i^q), \tag{12}$$

where k is non-negative and q is a finite power that determines the **s**hape (or curvature) of the ridge path through $p-$dimensional space, Goldstein and Smith (1974). The "ordinary" ridge estimators of equation (6) correspond to $q = 0$ in equation (12).

The 2-parameter family is quite versatile in the sense that most shrinkage paths considered in ridge regression literature are either special cases or limiting cases of this family. For example, $q = 1$ yields uniform shrinkage, $\delta_1 = \cdots = \delta_p$. In actual ridge practice, ties among eigenvalues are rare (except in designed experiments.) The common situation is $\lambda_1 > \cdots > \lambda_r > 0$ where $r = rank(X) \leq p$, and the first r shrinkage factors are then all unequal as long as $m > 0$, $m < r$ and $q \neq 1$ in equation

(12). Note that $q > 1$ would focus initial shrinkage upon major principal axes, $\delta_1 < \cdots < \delta_r$, while $q < 1$ focuses initial shrinkage along minor axes, $\delta_1 > \cdots > \delta_r$. These $q < 1$ (declining deltas) shrinkage patterns, when favored by the y data, have much greater potential for reduction in MSE risk via variance-bias trade-offs than do the $q > 1$ patterns.

The limit as q approaches $+\infty$ is optimal for the Gibbons(1981) "unfavorable case" where the true β vector lies along the eigenvector corresponding to the smallest regressor eigenvalue, λ_p. And the limit as q approaches $-\infty$ is essentially what Marquardt(1970) called "assigned-rank" regression. In both of these limiting cases, the shrinkage path travels along a series of "edges" of the principal-components regression hyper-rectangle. I have found that $q = \pm 5$ is usually adequate to roughly approximate these $q = \pm\infty$ limiting cases.

It is easily shown that the unknown extent of shrinkage that minimizes the MSE risk of $\delta_i \cdot c_i$ as a *linear* estimate of γ_i is

$$\delta_i^{MSE} = \frac{\gamma_i^2}{\gamma_i^2 + (\sigma^2/\lambda_i)} = \frac{\lambda_i}{\lambda_i + (\sigma^2/\gamma_i^2)} \ . \tag{13}$$

These equations can be solved to express γ_i and σ as functions of λ_i and of any trial value for the δ_i shrinkage factor to yield

$$\gamma_i = \pm\sigma\sqrt{\delta_i/[\lambda_i(1-\delta_i)]} = \pm\sigma/\sqrt{k\lambda_i^q} \ , \tag{14}$$

where the last expression follows only for ridge estimators in the 2-parameter family of equation (12). The likelihood that any given ridge estimation minimizes MSE risk is then defined by maximizing, by choice of the $\hat{\sigma}$ estimate and the $\pm$ signs, the likelihood that γ is of the form given in equation (14). The resulting closed-form solution, Obenchain(1981), is

$$\hat{k} = \hat{k}(q) = [\sum \lambda_j^{(1-q)}] \cdot \frac{[1 - R^2 \cdot CRL^2(q)]}{n \cdot R^2 \cdot CRL^2(q)}, \tag{15}$$

where the "*c*urlicue" function is

$$CRL(q) = \frac{\sum \left| r_j^o \right| \lambda_j^{(1-q)/2}}{\sqrt{\sum r_j^{o2} \sum \lambda_j^{(1-q)}}}. \tag{16}$$

Furthermore, the most likely $q-$shape is the one that maximizes $CRL(q)$, Obenchain(1975), and can be found by numerical search. Note that these maximum likelihood ridge estimators are more versatile than principal components regression in the sense that they use both the r_i^o and the $\lambda_i^{1/2}$ to select shrinkage factors anywhere in the range $0 \leq \delta_i < 1$.

A reasonable way to plot **traces** for the family of equation (12) is first to decide which p quantities will be plotted vertically, then to fix the value of the shape parameter q, and finally to plot with m of equation (11) on the horizontal axis over the range from $m = 0(k = 0)$ to $m = r(k = +\infty)$. In the Stata ADO files, numerical values for the k parameter are determined implicitly given values for m and q. And

all trace displays use m of equation (11) on the horizontal axis [instead of k] so that traces will not only have finite width but will also be much more easily comparable for different choices of $q-$shape.

1.3 The Stata ADO Files

This paper presents five Stata programs to estimate and evaluate models using maximum likelihood ridge regression.

RXrcrlq: Procedure **rxrcrlq** determines which $q-$shape (curvature) for the shrinkage path is most likely to contain the MSE-optimal ridge estimator. **rxrcrlq** searches a user-specified lattice of $q-$shapes within the range $-5 \leq q \leq +5$.

RXridge: Procedure **rxridge** performs generalized ridge calculations and displays 5 types of **traces** of specified $q-$shape: (i) shrunken coefficients, β_i^* of equation (10), (ii) estimated scaled (or relative) MSE, (iii) excess MSE eigenvalues [OLS minus ridge], (iv) inferior-direction cosines and (v) ridge shrinkage $\delta-$factors.

RXrmaxl: Procedure **rxrmaxl** computes 3 types of likelihood criteria to determine an ideal extent of shrinkage along a path of given $q-$shape: (i) classical, (ii) random-coefficients, and (iii) empirical Bayes.

RXrrisk: For specified true model parameters γ and σ and specified path $q-$shape, procedure **rxrrisk** computes and displays 5 types of **traces**: (i) expected coef-

ficients, (ii) true, scaled MSE, (iii) true excess MSE eigenvalues, (iv) the true inferior direction and (v) ridge shrinkage $\delta-$factors.

RXrsimu: For given model parameters and specified path $q-$shape, generate pseudo-random responses and a TRACE of the resulting true scaled, squared-error losses from shrinkage.

Note that the first 3 programs (RXrcrlq, RXridge and RXrmaxl) are the ones you should find most useful for data analysis and statistical inference. However, you may use the last 2 programs (RXrrisk and RXrsimu) to convince yourself that the estimation methods incorporated in the first 3 programs can be expected to work well in actual practice; in fact, this is the approach that we will use here in Sections 2, 3 and 4.

Our exploration of the Stata programs for likelihood-based ridge regression is organized as follows. Sections 2 and 3 show how true expected risks and simulated losses, respectively, can be expected to vary upon shrinkage when the regression parameters, β and σ, have known values. Unfortunately, these two preliminary sections have little to do with the usual analysis/inference situation in which ridge regression is actually applied, i.e. when the unknown regression parameters are to be estimated from an observed response variable (conditional on given predictor variables.) On the other hand, we will see in Section 4 that traces of maximum likelihood estimates of unknown, true MSE risks can mimic the most important features of their population

analogs from Sections 2 and 3. And illustrating this phenomenon certainly enhances the credibility of our graphical/likelihood approach to ridge regression analysis.

1.4 Why This Numerical Example?

The remainder of this paper uses the *Portland Cement* data of Hald (1952) to illustrate use of the five Stata programs. This dataset is well known and also quite small (only $n = 13$ observations on $p = r = 4$ predictor variables named **v3ca, v3cs, v4caf** and **v2cs.)** This is the numerical example I used in Obenchain(1984) to illustrate that one's data can suggest use of a relatively extreme path shape; here, our motivation for using the $q = -5$ path shape will be postponed until Section 4, below.

2 Known Model Parameters: RXRRISK

Let us assume that the true values of the uncorrelated components of β are $\gamma = G'\beta = (.646, .0, .323, .108)'$ and the true error standard deviation is $\sigma = .215$. [These are numerical values fairly close to their least squares estimates for the **heat** response variable of Hald(1952); namely, $\hat{\gamma} = (.657, .008, .303, .388)'$ and $\hat{\sigma} = .163$.] Figures 1 to 4 display **expected** shrinkage results for the $q = -5$ family of (12) using program **rxrrisk** for shrinkage risk analysis.

matrix rxsigma = (.215)

matrix rxgamma = (.646,.0,.323,.108)

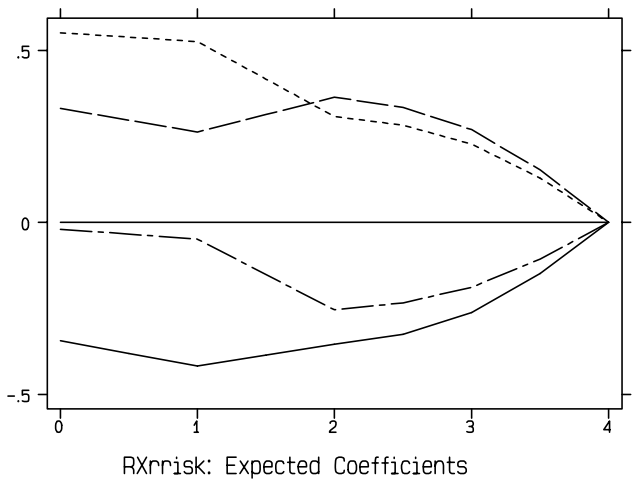

Figure 1: Expected Shrinkage Coefficients ($q = -5$)

rxrrisk heat v3ca v3cs v4caf v2cs, q(-5) m(2) t(0.001)

Figure 1 shows how the expected values of the ridge coefficients change as bias is introduced via the $q = -5$ family. The true regression coefficient vector, $\beta = G\gamma = (.552, .332, -.020, -.344)'$, is displayed at $m = 0$ in Figure 1. Thus the first two, true coefficients are positive while the last two are negative. In particular, note that bias introduced by shrinkage along the $q = -5$ path can tend to make β_3 somewhat more negative than its true value of $-.020$.

Figure 2 gives the associated scaled (or "relative") MSE risks defined as follows. Risk is expected loss (or "mean" squared error), and the scaled risk values plotted in Figure 2 are the diagonal elements of the mean squared error matrix each divided

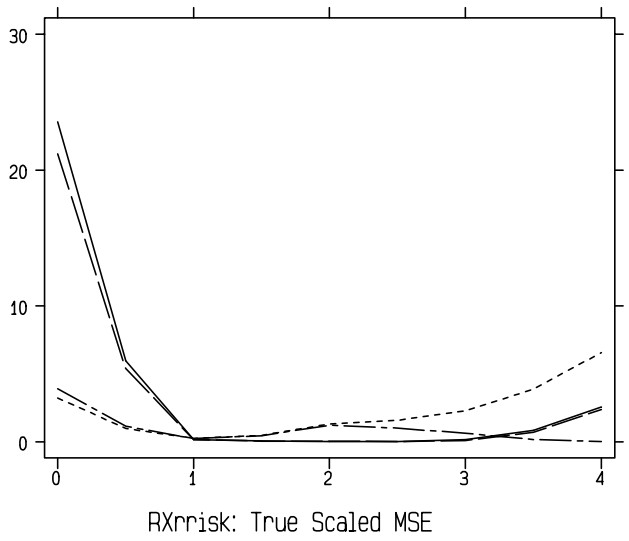

Figure 2: True Scaled MSE Risks ($q = -5$)

by σ^2. Scaled risk values measure the uncertainty in an estimate as a multiple of the variance of a single observation. Scaled risk values also have the advantage of being **known** values for least-squares estimates even when regression parameters are unknown. For example, the values at the left extreme ($m = 0$) of Figure 2 are the diagonal elements of $(X'X)^{-1}$, which do not depend upon $\beta = G\gamma$ or σ.

The eigenvalues of the difference in scaled risk matrices (least squares minus ridge) are displayed in Figure 3. As long as these eigenvalues are all non-negative, no linear combination of least squares coefficients has smaller risk than the corresponding linear combination of ridge coefficients.

At most one excess eigenvalue can be negative, Obenchain(1978), and the cor-

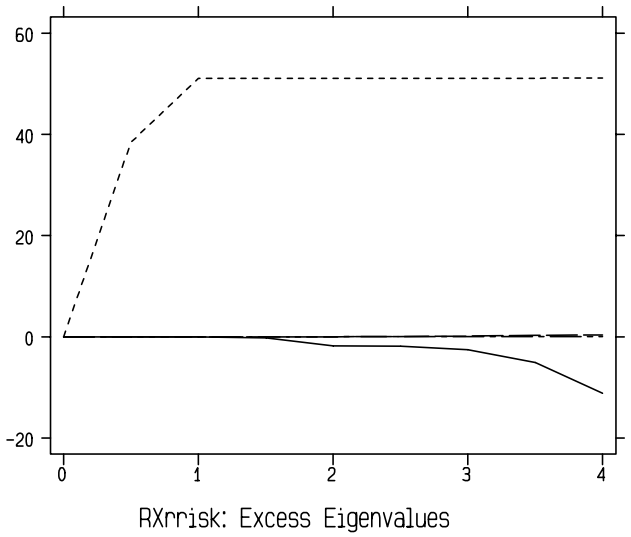

Figure 3: True Excess Eigenvalues (Least Squares minus Ridge)

responding normalized eigenvector, Figure 4, points in the **inferior-direction** of $p-$dimensional space along which ridge has higher risk than least squares. For example, the 1st and 3rd direction cosines (for variables **v3ca** and **v4caf**) are nearly equal when an inferior direction appears at $m = 1.5$ in Figure 4. Linear combinations like $\beta_1 - \beta_3$ are thus essentially orthogonal to the inferior direction at this m, so the ridge estimate of $\beta_1 - \beta_3$ probably has lower risk than least squares. But the ridge estimate of $\beta_1 + \beta_3$ near $m = 1.5$ can possibly have higher risk than least squares because this linear combination has a relatively large projection onto the inferior direction.

Figures 1 to 4 indicate that our numerical example is amenable to ridge shrinkage with $q = -5$ in equation (12). The trace of the scaled risk matrix decreases from 51.8

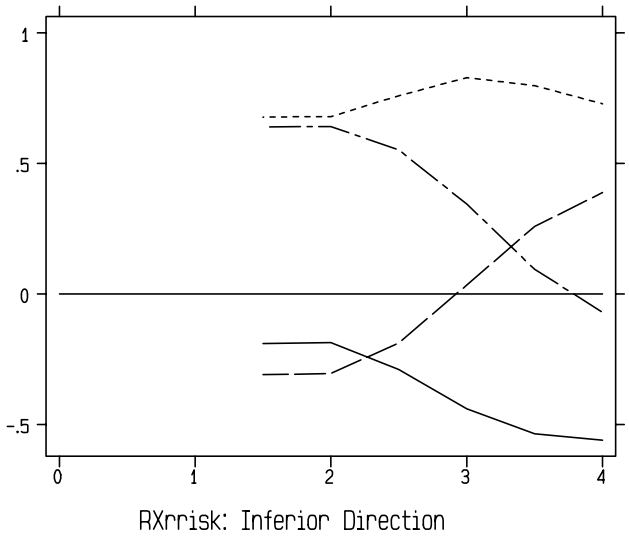

Figure 4: True Inferior Direction Cosines ($q = -5$)

at $m = 0$ to 0.787 at $m = 1.0$ and then starts increasing again. Thus an m value of about 1 is risk optimal when $q = -5$, and this is like saying that ill-conditioning has effectively reduced the rank of the regressor matrix by 1 from 4 to 3. This makes very good sense in this example because the regressor matrix comes from a "mixture" equipment with the four regressors adding (except for a relatively large round-off error) to 100%.

3 Simulated Responses: RXRSIMU

The logical next step in exploring our numerical example is to use pseudo-random numbers to simulate a response vector for this model using procedure **rxrsimu**. The

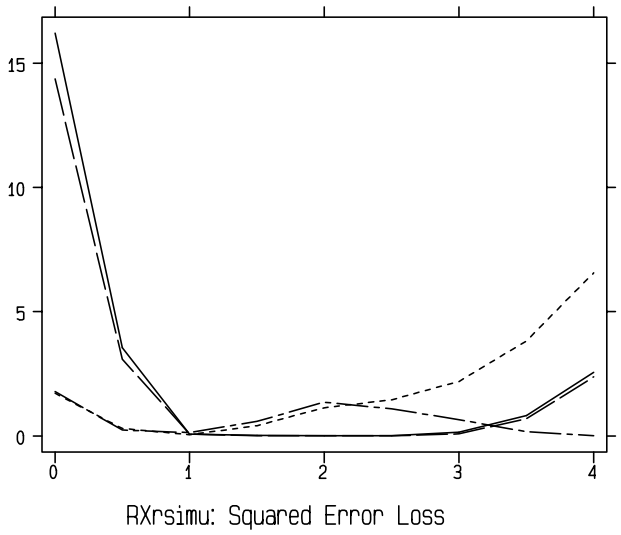

Figure 5: Scaled, Squared-Error Losses ($q = -5$)

results from a single invocation of **rxrsimu** are given in Table I and Figure 5.

matrix rxsigma = (.215)

matrix rxgamma = (.646,.0,.323,.108)

rxrsimu heat v3ca v3cs v4caf v2cs, q(-5) m(2) t(0.001)

Table I. RXrsimu Responses and Expected Values

	ysim	yexp		ysim	yexp
1	-1.123823	-1.115518	8	-1.119887	-1.29178
2	-1.251552	-1.477537	9	-.2948058	-.2422573
3	.6228011	.7167597	10	1.434455	1.351851
4	-.4981516	-.3721892	11	-1.18242	-.8970727
5	-.0099682	-.0052348	12	1.404445	1.091648
6	.42115	.6511954	13	1.229661	1.043567
7	.3680945	.5465683			

Note that minimum overall loss occurs at about $m = 1.0$ in Figure 5. Also, remember that the expected value of the scaled, squared-error loss trace of Figure 5 would be the scaled MSE risk trace of Figure 2.

4 Data Analysis/Inference: RXRIDGE

Let us now continue our numerical example from Sections 2 and 3 by analyzing the simulated responses using procedure **rxridge** as if we had no knowledge of the true regression parameter settings used to generate the data.

rxridge ysim v3ca v3cs v4caf v2cs, q(-5) m(2) t(0.001)

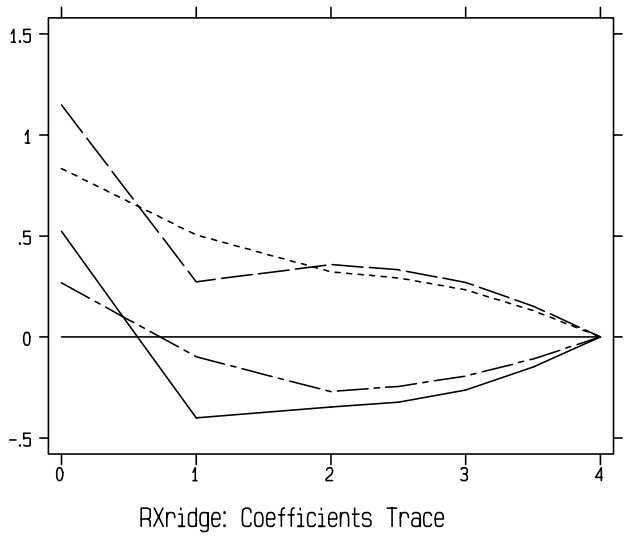

Figure 6: Trace of Shrinkage Coefficients ($q = -5$)

RXridge: Shrinkage Path has Qshape =-5.00

RXridge: Estimated Sigma = .21311153

RXridge: Uncorrelated Components... Number of obs = 13

	Coefficient	Std. Error	t-statistic	p-value	Lower 95%	Upper 95%
c1	.6526046	.0411443	15.861	0.000	.5577258	.7474834
c2	-.022923	.0490037	-0.468	0.652	-.1359258	.0900798
c3	.2709772	.1424142	1.903	0.094	-.0574305	.599385
c4	1.364187	1.526713	0.894	0.398	-2.156419	4.884793

Note that only the estimates of the 1st and possibly 3rd uncorrelated components are statistically significant, but the 4th component is huge, numerically. Thus our

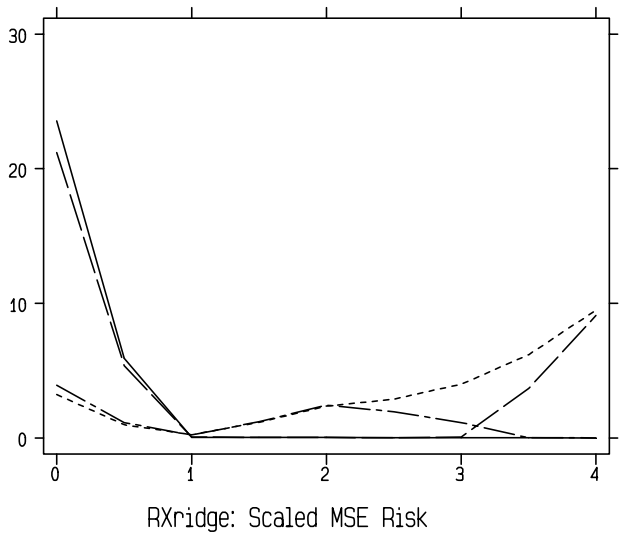

Figure 7: Estimated Scaled MSE ($q = -5$)

rxrsimu generated response vector is even more susceptible to ill-conditioning in X than were the original **heat** data of Hald(1952).

Note in Figure 6 that all four least-squares estimates for coefficients are positive (m=0) and that shrinkage to at least $m = 1$ is required to produce ridge coefficients with the "right" numerical signs.

Figures 7-9 are traces of **estimates** of scaled risks, excess eigenvalues, and inferior direction cosines, respectively (Obenchain 1978, 1981). These traces are all based upon normal-theory maximum-likelihood, but scaled risk estimates have, first, been adjusted using known constants that make them unbiased and then truncated, if necessary, so as to have correct range (i.e. no scaled risk estimate is given that is

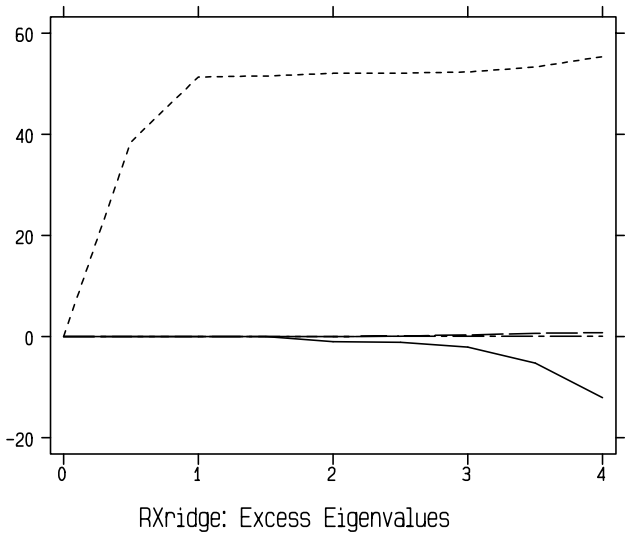

Figure 8: Estimated Excess SMSE Eigenvalues ($q = -5$)

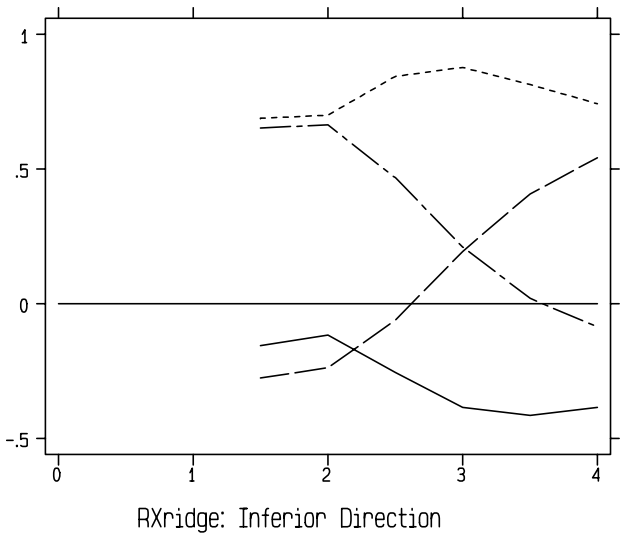

Figure 9: Estimated Inferior Direction Cosines ($q = -5$)

below its scaled variance lower limit.)

Visual examination of Figures 6-9 suggests that the ridge solution at $m = 1.0$ in the $q = -5$ family has much more desirable risk characteristics than does the least squares solution.

5 Shrinkage Path Shape: RXRCRLQ

So far we have plotted traces using only shape $q = -5$ in (12). This is because $q = -5$ was the MSE optimal shrinkage path shape for the original Hald(1952) data. And we can use procedure **rxrcrlq** to verify that this is also the best choice for the **rxrsimu** generated data.

rxrcrlq ysim v3ca v3cs v4caf v2cs

Table II. RXrcrlq Choice of Shrinkage Path q-shape

q-shape	**M**CAL	**k**	**C**RL(q)	**C**hi-Square
+5	3.9620787	1.760e+08	.05583205	45.426121
+4	3.961938	3416311.2	.05593933	45.425969
+3	3.9597236	62767.277	.0576044	45.423578
+2	3.8986512	460.94542	.09406287	45.353457
+1	1.4705107	.58134686	.59741232	39.943159
0.0	1.7167549	2.4715085	.78595994	33.585201
-1	2.0447908	40.498435	.83293916	30.942830
-2	2.1182805	727.90415	.86989888	28.259285
-3	2.1182927	13417.735	.90110205	25.337028
-4	2.1129101	254461.68	.92596664	22.318394
-5	2.1115136	4990888.1	**.94490765**	**19.354315**

RXrcrlq uses the maximum likelihood equations (15) and (16) to evaluate alternative $q-$shapes for the shrinkage path. Here, the most likely $q-$shape is -5 because it achieves maximum CRL(q) and minimum Chi-Squared in Table II. This minimum Chi-Squared has degrees-of-freedom= 2 and significance level= 0.000. Thus the 2-parameter generalized ridge family is probably too restrictive (unlikely to contain the MSE optimal shrinkage factors) for this example.

Is there a family of shrinkage (δ) factors "less restrictive" than equation (12) that we could consider? Not really; I know of no proposals for, say, a 3−parameter family. A full $r-$parameter solution (in which each δ factor is estimated separately) is possible, but this would impose no "smoothness" requirements, whatsoever, on shrinkage factors. (In the current example, the data strongly suggest taking δ_2 much smaller than δ_1 or δ_3, but there is very little potential for MSE reduction by such a "greedy" tactic because $\hat{\gamma}_2 = -.02$ is already tiny, numerically.) Besides, $r-$parameter estimates are not amenable for visual display using ridge **traces**!

It's a straightforward task to generate traces for several different values of q and to make a choice (either objective or subjective) of which shape one likes best. These traces can change shape and, thus, interpretation quite drastically as q changes. Obviously unfavorable choices of q will have minimum SMSE risk either at or very close to $m = 0$ in Figures 2 and 7. Furthermore, a negative excess eigenvalue will not only appear for very small m values in Figures 3 and 8 when q is unfavorable, but this negative eigenvalue will also dominate the most positive eigenvalue in absolute magnitude. Anyway, the most likely $q-$shape (which is -5 in our current example) usually strikes me as being adequately "general" even when the **rxrcrlq** chi-squared statistic is significantly greater than zero.

A search on a finer lattice of q values over a wider range than $-5 \leq q \leq +5$ could be considered, of course, but we must remember that the primary purpose of the **rxrcrlq**

calculations of Table II is simply to interest us in examining the corresponding **trace** displays.

6 Shrinkage Extent: RXRMAXL

Other maximum likelihood approaches besides the classical, fixed coefficient approach of Obenchain (1975, 1981) are possible, but they do not yield closed form expressions for the optimal k or m given q. The empirical Bayes approach of Efron and Morris (1977) and the random coefficient method of Golub, Heath, and Wahba (1979) and Shumway (1982) are two maximum likelihood alternatives implemented in **rxrmaxl**. This program "monitors" all three of the above likelihood criteria on a lattice of m values, referring to them as CLIK, EBAY, and RCOF, respectively.

Table III. RXrmaxl Choice of Shrinkage Extent

MCAL	CLIK		EBAY		RCOF	
0	∞		∞		∞	
0.5	1.742e+12		83.985422		84.219263	
1	9.889e+11		74.129038		74.567306	
1.5	445435.87		31.003878		32.064207	
2	738.64505		19.480671	<<	20.780788	<<
2.5	27.939063	<<	28.330190		21.782698	
3	35.164964		71.501773		31.229391	
3.5	39.875945		151.47765		39.478093	
4	45.465477		256.22102		45.465477	

The maximum-likelihood choices for m-extent of shrinkage are the ones that minimize the CLIK, EBAY or RCOF criteria, above. I used **rxrmaxl** to produce Table III on the **ysim** variable generated by **rxrsimu** within the $q = -5$ family. (CLIK is a minus two log likelihood ratio whose minimum has an asymptotic chi-squared distribution with two degrees-of-freedom, as in Table II. EBAY and RCOF are normalized so that they cannot become negative, but their minima do not apparently have asymptotic chi-squared distributions.) Anyway, the EBAY and RCOF criteria both favor $m = 2$ when $q = -5$ while CLIK favors $m = 2.5$. Thus, using the "$2/r$-ths

Rule-of-Thumb" of Obenchain(1978) for the extent of shrinkage likely to be "good" (i.e. to dominate least-squares in a **matrix** MSE sense) in this $p = r = 4$ predictor model, we again find that shrinkage to at least $m = 2 \times 2/r = 1$ is highly desirable.

The primary reservation that comes to my mind concerning the calculations of Table III is that they are difficult to plot, at least simultaneously. All start at $+\infty$ at $m = 0$, and EBAY can also be very large as m approaches p. And, again, minimum values are not comparable. However difficult they may be to produce, plots of these minus two log likelihoods are of interest because one needs to "see" how "flat" each is near its minimum.

7 Numerical Example: SUMMARY

The most striking feature of our example is the extent to which the **rxridge** estimates of Section 4 mimic their expected values from **rxrrisk** of Section 2. I assure the reader that this mimicry is typical rather than an artifact of the single set of simulated responses generated using **rxrsimu** in Section 3. (The overall signal-to-noise ratio here was $\beta'\beta/\sigma^2 = 2.5$ when the diagonal elements of $X'X$ were scaled to equal $(n-1) = 12$, and one might expect much less mimicry with much lower ratios.) To any skeptic, I simply suggest... "Why not try it for yourself?"

Next, I ask you to reexamine the estimated traces of Section 4. Isn't it remarkable how much incredibly detailed information is contained in these traces concerning the

extent and effects of shrinkage on ill-conditioned estimates of regression coefficients?

8 Stata Syntax

rxrcrlq [depvar [varlist] [if exp] [in range]] [, qmin(#) qmax(#) nq(#) rescale(#) tol(#)]

rxridge [depvar [varlist] [if exp] [in range]] [, msteps(#) qshape(#) rescale(#) tol(#)]

rxrmaxl [depvar [varlist] [if exp] [in range]] [, msteps(#) qshape(#) omdmin(#) rescale(#) tol(#)]

matrix rxsigma = (#)

matrix rxgamma = (#,...,#)

rxrrisk [depvar [varlist] [if exp] [in range]] [, msteps(#) qshape(#) omdmin(#) rescale(#) tol(#)]

rxrsimu [depvar [varlist] [if exp] [in range]] [, msteps(#) qshape(#) start(#) rescale(#) tol(#)]

qmin(#) specifies the minimum $q-$shape to evaluate. The default value is **qmin**$=-5$, and **qmin** cannot be reset to any value greater than -2.

qmax(#) specifies the maximum $q-$shape to evaluate. The default value is **qmax**= +5, and **qmax** cannot be reset to any value less than +2.

nq(#) specifies an integer number of $q-$shape values to evaluate between **qmin** and **qmax**, inclusive. The default value is **nq**= 21, and **nq** cannot be reset to any integer value less than 9.

msteps(#) specifies the number of steps per unit increase in m, the multicollinearity allowance parameter; the default value is 4. The total number of steps along the generalized shrinkage path from the Least Squares solution ($m = 0$) to all ZERO coefficients ($m = r$) will thus be 1+(**msteps**$\times r$), where $r = rank(X)$.

qshape(#) controls the shape (or curvature) of the generalized shrinkage path through likelihood space; the default value is 0, which yields Hoerl-Kennard "ordinary" ridge regression. **qshape**=1 yields uniform shrinkage, and all **qshape** values between +5.0 and −5.0 are allowed.

rescale(#) controls the scaling of the response variable and all p predictor variables in the varlist. The default value is **rescale**= 1 to scale all "centered" variables to have sample variance 1 (sample sum-of-squares equal to one fewer than the number of observations.) To retain the scaling of variables given in the Stata ".DTA" file, specify **rescale(0)**.

tol(#) specifies the search convergence criterion and defaults to 0.01.

omdmin(#) is the strictly positive minimum value to be used for calculation of $(1-\delta)$ shrinkage factors; the default is **omdmin**$= 10e-13$.

start(#) controls Stata's uniform() random number seed value, and the RXrsimu default value is 12345. If you make repeated **rxrsimu** runs without changing this seed, you will get the same pseudo-random values each time! When you do change **start**, make it positive.

Restrictions The RXridge programs (rxrcrlq, rxridge, rxrmaxl, rxrrisk, and rxrsimu) do not treat multiple regression models that lack an intercept (constant) term; the number, p, of (nonconstant) predictor variables, X, in the varlist must be at least 2; if p is greater than 20, Stata will refuse to draw **trace** plots; the depvar, y, must be nonconstant; and no missing values are allowed. RXridge programs internally "center" all variables to have mean zero, and the fitted (hyper)plane then always passes through $\bar{y} = 0$ at $\bar{X} = 0$. The implied $y-$intercept at the original X origin can, of course, be determined implicitly as the coefficients for the p, nonconstant regressors change, but this $y-$intercept is not calculated by the RXridge programs.

In addition to Stata, I have programmed my maximum-likelihood ridge algorithms in SAS/IML, S-PLUS and GAUSS. And Bernhard Walter, Technische Universitaet

Muenchen, has created splendidly interactive routines for XLisp-Stat. However, the most complete implementation of my algorithms is undoubtedly provided by my stand-alone systems for MS-DOS personal computers: RXridge.EXE, RXtraces.EXE and PathProj.EXE, Obenchain(1991) and Nash(1992). For example, RXridge.EXE calculates inference intervals (classical, confidence and Bayes HPD) for shrunken coefficients as well as performs ridge residual analyses (outlying responses and/or high leverage regressor combinations.) I distribute all of the above software systems as freeware.

References

Efron, B. and Morris, C. (1977). Comment on "A simulation study of alternatives to ordinary least squares," by A. P. Dempster, Martin Schatzoff, and Nanny Wermuth. **Journal of the American Statistical Association** 72, 91-93.

Gibbons, D. G. (1981). "A simulation study of some ridge estimators." **Journal of the American Statistical Association** 76, 131-139.

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

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.

Hald, A. (1952). **Statistical Theory and Engineering Applications.** New York:John Wiley. (Portland Cement Data, page 647).

Hoerl, A. E. (1962). "Applications 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 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. (errata: 723.)

Nash, J. C. (1992). "Statistical shareware: illustrations from regression techniques." **The American Statistician** 46, 312-318.

Obenchain, R. L. (1975). "Ridge analysis following a preliminary test of the shrunken hypothesis." **Technometrics** 17, 431-441.

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

Obenchain, R. L. (1981). "Maximum likelihood ridge regression and the shrinkage pattern alternatives." Unpublished review; 74 pages. (**I.M.S. Bulletin** 10, 37; Abstract 81t-23.)

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.)

Obenchain, R. L. (1991). "Ridge regression systems for MS-DOS personal computers." **The American Statistician** 45, 245-246.

Obenchain, R. L. and Vinod, H. D. (1974). "Estimates of partial derivatives from ridge regression on ill-conditioned data." **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.

Shumway, R. H. (1982). "Maximum likelihood estimation of the ridge parameter in linear regression." Technical Report, Division of Statistics, University of California at Davis.

Vinod, H. D. (1976). "Application of new ridge methods to a study of Bell System scale economies." **Journal of the American Statistical Association** 71, 835-841.