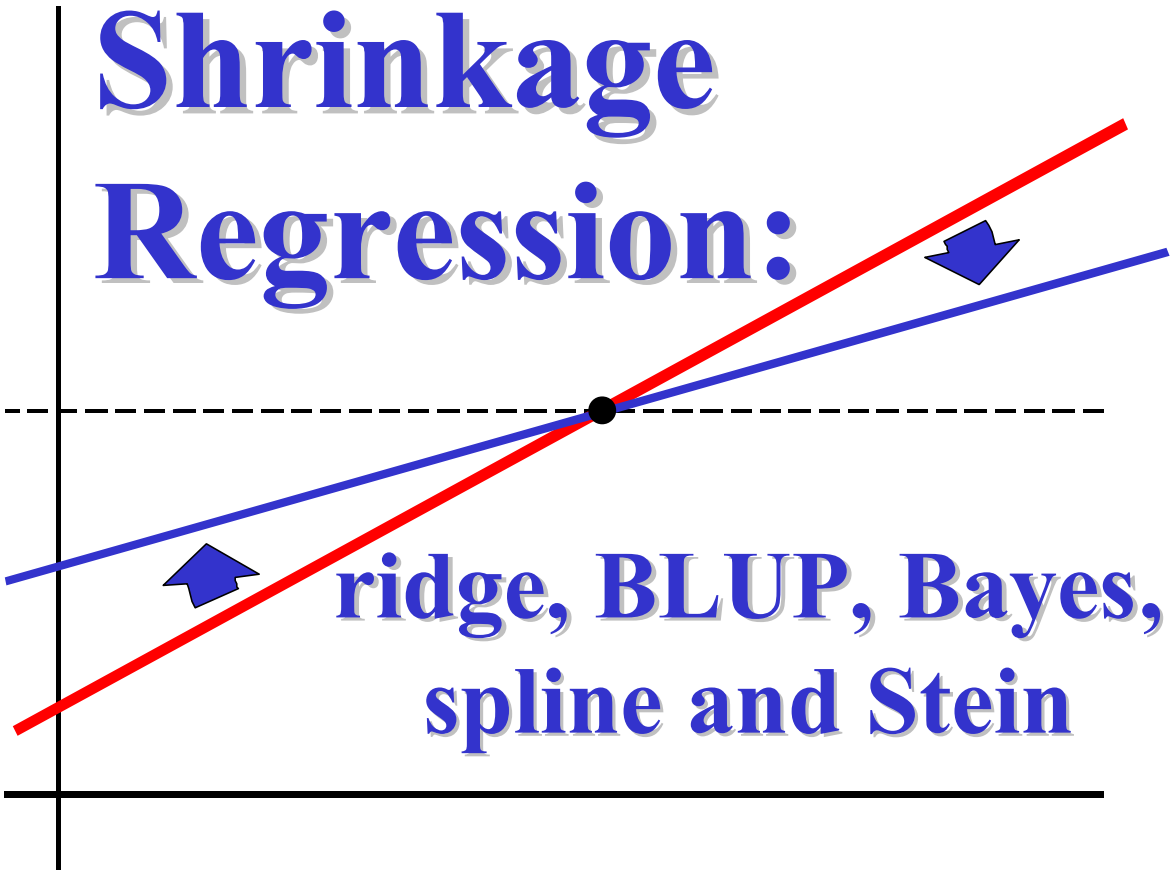


Chapter 06: Risk Estimation & Simulation

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 $\nu = N - R - 1$.

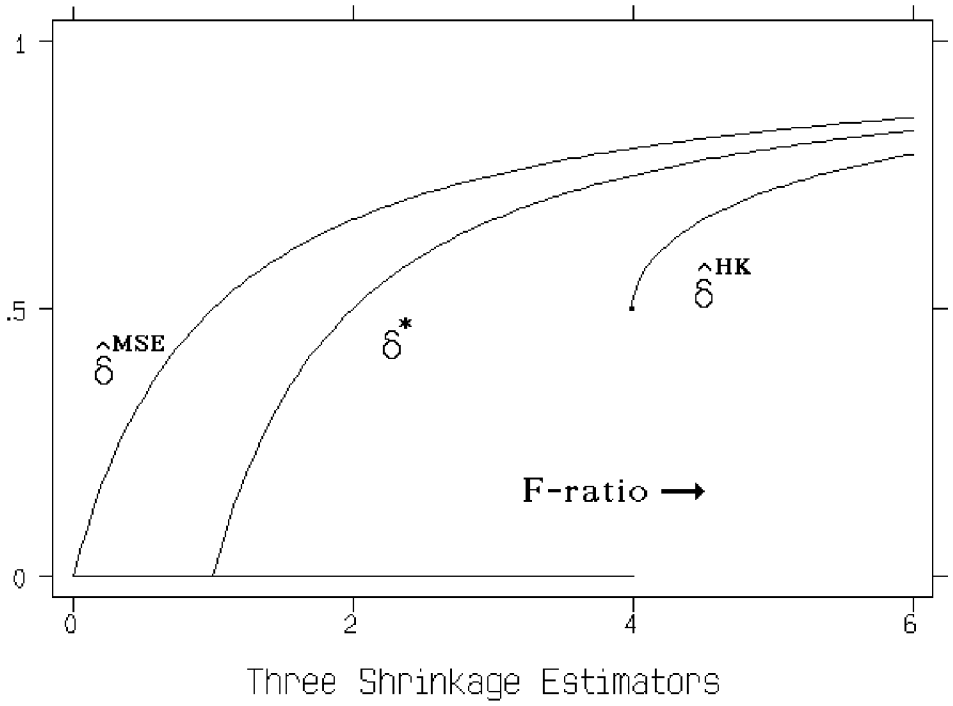
For comparison with { 6.20 } and { 6.21 } , we remark that Hemmerle(1975) showed that the heuristic "fixed-point" iteration of Hoerl and Kennard(1970a,b) converges to the almost drastic shrinkage value:

$$\begin{aligned} \widehat{\delta}_i^{HK} &= 0 && \text{if } 0 \leq F_i \leq 4, \\ &= (1 + \sqrt{1 - 4 \cdot F_i^{-1}}) / 2 && \text{otherwise.} \end{aligned} \qquad \{ 6.22 \}$$

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

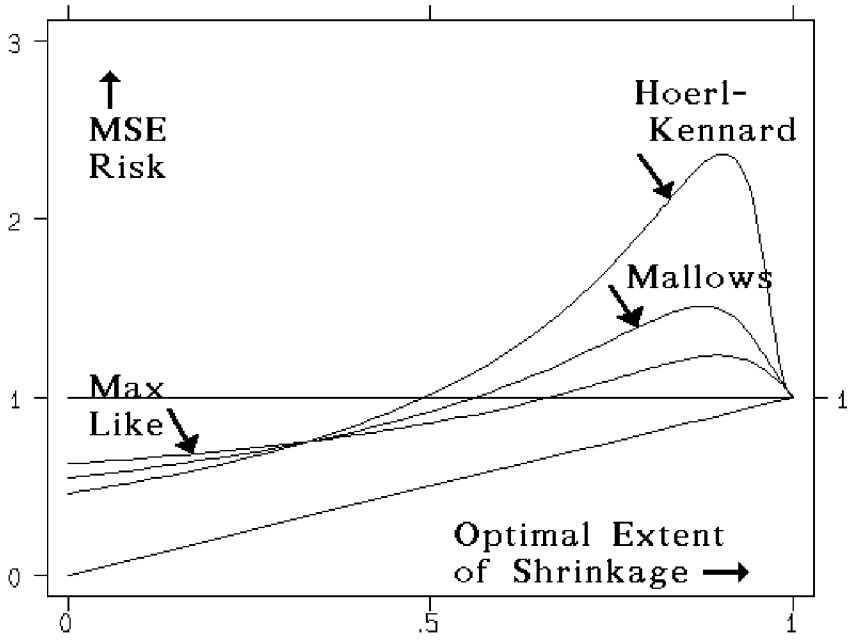
Figures 6.2, 6.3 and 6.4 display simulated mean-squared-error risk profiles for the three fixed coefficient shrinkage estimators whose relative extents of shrinkage were displayed in Figure 6.1. These estimators are: maximum likelihood shrinkage as in { 6.22 }, Mallows' minimum estimated-risk from { 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...**

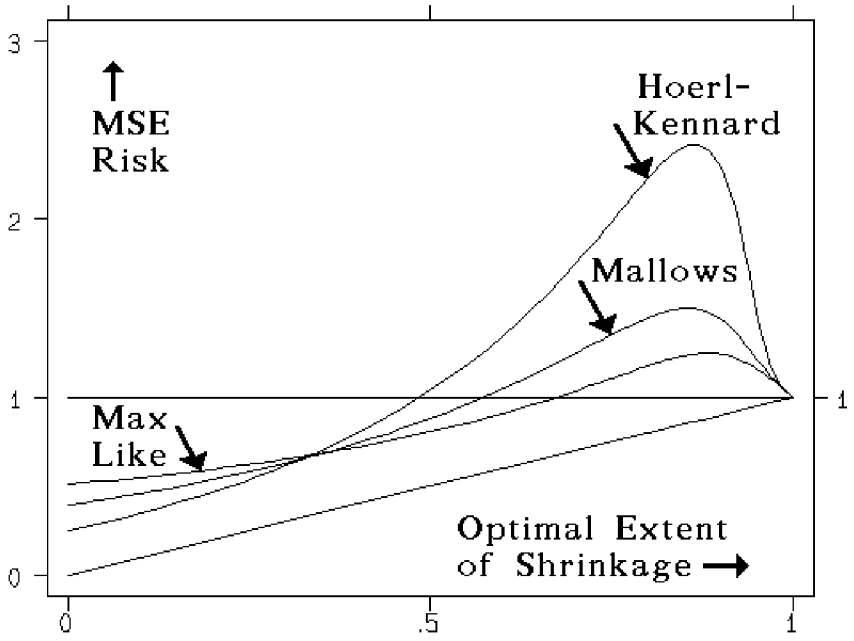
Figure 6.2 Simulated Fixed Coefficient Risk when Error Degrees-of-Freedom = 1.



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

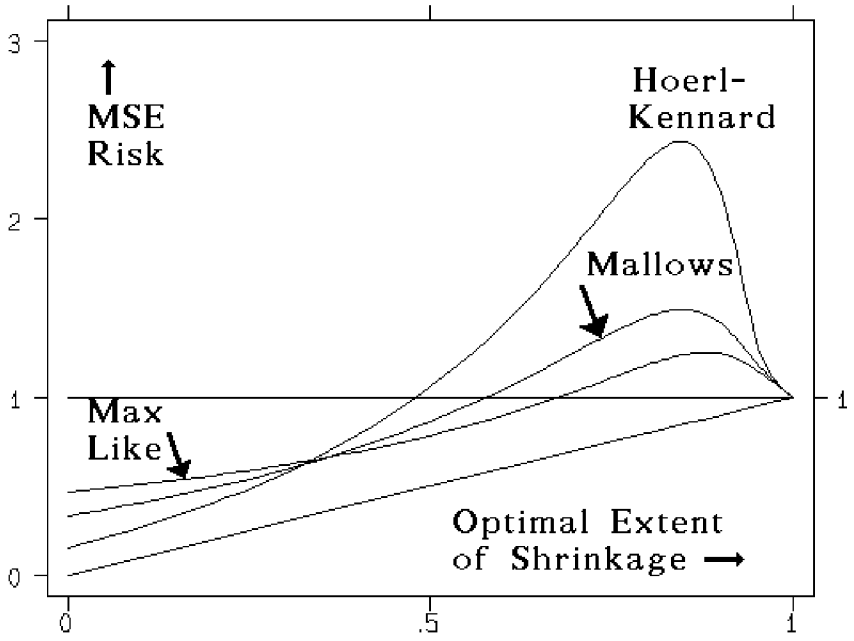
Figure 6.3 Simulated Fixed Coefficient Risk when Error Degrees-of-Freedom = 5.



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...

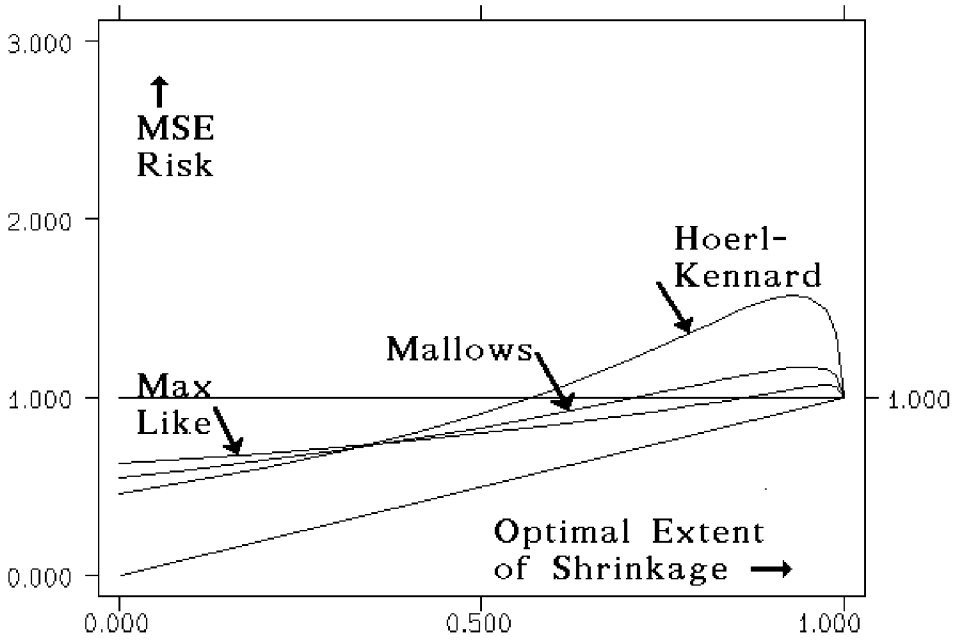
Shrinkage estimation does a much better job of either reducing and/or limiting increases in MSE risk when coefficients vary randomly than when they are fixed.

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%.

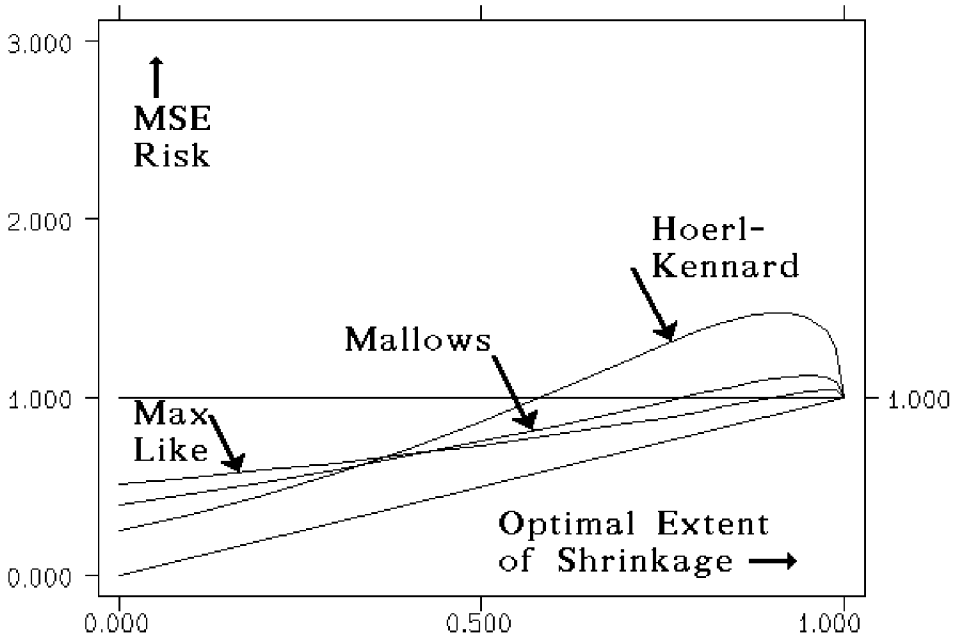
Figure 6.5 Simulated Random Coefficient Risk when Error Degrees-of-Freedom = 1.



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

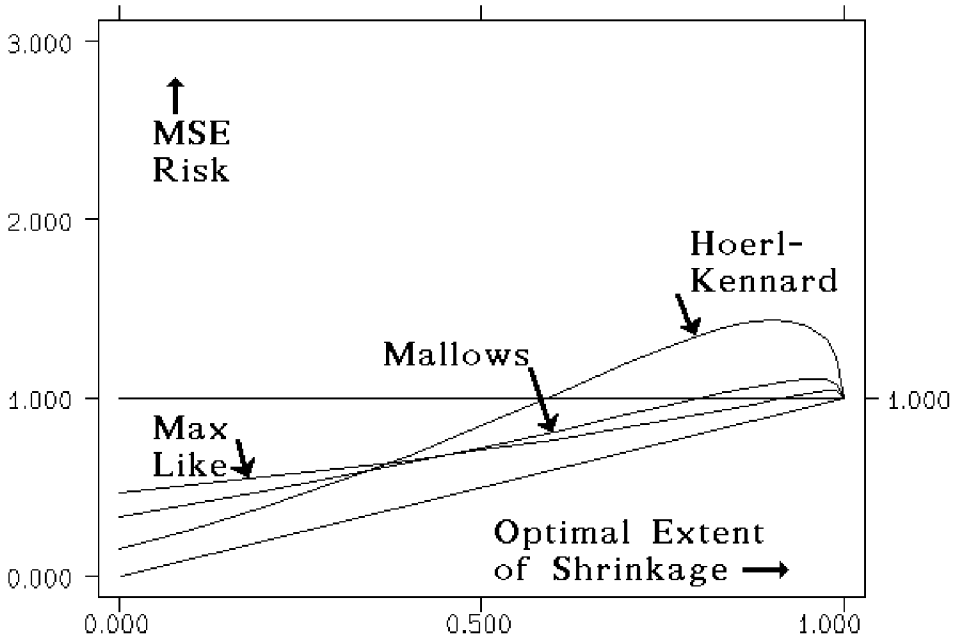
Figure 6.6 Simulated Random Coefficient Risk when Error Degrees-of-Freedom = 5.



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.