



Chapter 05: Normal Theory Maximum Likelihood

Bob Obenchain, Ph.D.
softRx freeware
13212 Griffin Run
Carmel, Indiana 46033-8835

Copyright © 1985-2004 Software Prescriptions

Chapter 5: NORMAL THEORY

MAXIMUM LIKELIHOOD

Here in Chapter 5 we discuss Normal-theory methods of identifying which shrinkage regression coefficient estimates are most likely to have optimal mean-squared-error, i.e. which set of shrinkage factors are most likely to be the target values, δ_i^{MSE} , of equation { 4.6 } for $1 \leq i \leq R$. Of course, once you have located the extent of shrinkage most-likely-to-be-optimal, you still have the option of limiting shrinkage using the “(2/R)ths Rule-of-Thumb” (defined in Section §4.2) and thus maximizing (in some weaker sense) the likelihood of ending up with an estimator that has good multivariate mean-squared-error characteristics.

The first three sections of Chapter 5 explore methods for classical (fixed-coefficient) linear models. Section §5.1 gives a brief review of the relatively well-known unrestricted maximum likelihood theory under which least-squares regression coefficients are BLUE (Best Linear Unbiased Estimates.) Section §5.2 starts our exploration of methods for statistical inference concerning shrinkage by first defining the “likelihood” that any given amount of shrinkage yields optimal mean-squared-error and by then deriving its maximum, as in Obenchain(1975). In Section §5.3, we develop a closed-form expression for the shrinkage estimator most likely to attain minimal mean-squared-error within the 2-parameter generalized shrinkage family, as in Obenchain(1981).

The last two sections of Chapter 5 explore models with random (stochastic) coefficients. Section §5.4 reviews Henderson's BLUP (Best Linear Unbiased Prediction) theory for “mixed” models (containing both fixed and random coefficients.) In practical applications, maximum likelihood estimates of mixed model coefficients usually end up being neither “linear” nor “unbiased”. After all, just as in purely fixed-coefficient formulations, optimal estimates for random-coefficient models depend upon unknown parameters which, when estimated from the data at hand, yield operational analogs that are non-linear and biased. Section §5.5 develops much more detailed results on maximum likelihood estimation for the special case of purely-random models with a single variance component, as in Golub, Heath, and Wahba(1979) and Shumway(1982).

5.1 Unrestricted Maximum Likelihood and BLUE Theory

As we saw in Chapter 2, the general linear model is commonly stated using the pair of vector/matrix equations: $E(y | X) = 1\mu + X\beta$ and $V(y | X) = \sigma^2 I$. These are equations { 2.1 } and { 2.2 }, i.e. before the response vector, y , and the regressor matrix, X , are “centered” by subtracting off column means. As before, β is the column vector of unknown regression coefficients, μ is the unknown y intercept (the expected response corresponding to a

null row of X), and σ^2 is the unknown residual variance. Under Normal-distribution-theory, the joint likelihood function for μ , β and σ^2 is then

$$L(\mu, \beta, \sigma^2) = (2\pi\sigma^2)^{-N/2} e^{-u^2/2\sigma^2}, \quad \{ 5.1 \}$$

where u^2 is the quadratic form

$$u^2 = (y - 1\mu - X\beta)^T (y - 1\mu - X\beta). \quad \{ 5.2 \}$$

Also as before, we will use \bar{y} and \bar{x}^T to represent the mean values of the original response values and regressor combinations, respectively. This allows us to decompose u^2 into a sum of two terms, thereby isolating the contributions to u^2 resulting from coordinates parallel to the 1 vector from those contributions from coordinates orthogonal to the 1 vector. Returning now to the convention that the y vector and the X matrix have been "centered", as in Section §2.1, equations { 2.3 } and { 2.4 }, we can rewrite equation { 5.2 } as

$$u^2 = (\bar{y} - \mu - \bar{x}^T\beta)^2 + (y - X\beta)^T (y - X\beta). \quad \{ 5.3 \}$$

For any estimate $\hat{\beta}$ of β , the first term in the above expression for u^2 can always be made to vanish simply by taking $\hat{\mu} = \bar{y} - \bar{x}^T\hat{\beta}$. As a result, we can drop μ from further, explicit consideration and view the likelihood as really being a function of only the β and σ^2 estimates.

It is well known that the global, unrestricted maximum of $L(\beta, \sigma^2)$ is then achieved at

$$L(b^0, \hat{\sigma}^2) = (2\pi e\hat{\sigma}^2)^{-N/2} \quad \{ 5.4 \}$$

where $b^0 = X^+y = Gc$ is the least squares estimator of β [as in equation { 2.6 }] and $N \cdot \hat{\sigma}^2 = y^Ty \cdot (1-R^2)$ is the minimum value for u^2 of equation { 5.3 }. In other words, the vector of ordinary least squares coefficients, b^0 , is the maximum-likelihood estimator of the unknown, true β vector under Normal-distribution theory. Note that the maximum-likelihood estimator of σ^2 , namely $\hat{\sigma}^2 = y^Ty \cdot (1-R^2) / N$, is larger than the usual unbiased estimator [s^2 of equation { 2.22 }] by a multiplicative factor of $[N / (N - R - 1)]$. Although these results are quite well known, we will now review a derivation of { 5.4 } to illustrate some of the techniques we will also use in Section §5.2.

Our initial step toward maximizing the Normal-theory likelihood of equation { 5.1 } will be to minimize $u^2 = (y - X\hat{\beta})^T (y - X\hat{\beta})$ by choice of $\hat{\beta}$ [and then set $\hat{\mu} = \bar{y} - \bar{x}^T\hat{\beta}$]. Thus note that the vector of partial derivatives of u^2 with respect to β is

$$\partial(u^2)/\partial\beta = 2 \cdot X^TX\beta - 2 \cdot X^Ty, \quad \{ 5.5 \}$$

while the matrix of second partial derivatives is non-negative definite

$$\partial^2(u^2)/\partial\beta^2 = 2 \cdot X^TX. \quad \{ 5.6 \}$$

Thus any solution to $\partial(u^2)/\partial\hat{\beta} = 0$ yields the minimum value for u^2 , and $\hat{\beta} = b^o = X^+y$ is always a solution to these so-called “normal equations,” $X^T X \hat{\beta} = X^T y$.

Our final step in maximizing the Normal-theory likelihood of equation { 5.1 } is to choose the estimator of σ^2 . Equivalently, we can minimize the minus-twice-log-likelihood, $-2 \cdot \ln(L) = N \cdot \ln(2\pi\sigma^2) + u^2/\sigma^2$. The first partial derivative is thus

$$\partial[-2 \cdot \ln(L)]/\partial(\hat{\sigma}^2) = \frac{N}{\hat{\sigma}^2} - \frac{u^2}{(\hat{\sigma}^2)^2}, \quad \{ 5.7 \}$$

while the second partial derivative is

$$\partial^2[-2 \cdot \ln(L)]/\partial(\hat{\sigma}^2)^2 = -1 \cdot \frac{N}{(\hat{\sigma}^2)^2} + 2 \cdot \frac{u^2}{(\hat{\sigma}^2)^3}. \quad \{ 5.8 \}$$

Note that $\partial[-2 \cdot \ln(L)]/\partial(\hat{\sigma}^2) = 0$ admits two solutions, $\hat{\sigma}^2 = u^2/N$ and $\hat{\sigma}^2 = +\infty$. But the infinite solution can be eliminated from consideration because the second partial derivative vanishes at this point. On the other hand, $\partial^2[-2 \cdot \ln(L)]/\partial(\hat{\sigma}^2)^2$ assumes the strictly positive value $[N/\hat{\sigma}^4]$ at $\hat{\sigma}^2 = u^2/N$ under the (almost certain) condition that the minimum sum-of-squares of residuals is strictly positive. This establishes that the maximum value for the likelihood of equation { 5.1 } is indeed given by { 5.4 }, where $\hat{\mu} = \bar{y} - \bar{x}^T b^o$, $b^o = X^+y$, and $\hat{\sigma}^2 = y^T y \cdot (1-R^2)/N$.

The least-squares estimator, b^o , is an unbiased estimator of β under the usual assumption that the given expectation equation, $E(y|X) = 1\mu + X\beta$, is correct. Similarly, under the usual assumption that the dispersion equation, $V(y|X) = \sigma^2 I$, is correct, b^o is also the minimum variance estimator within the class of linear, unbiased estimators of β . As a result, b^o is commonly said to be the BLUE (Best Linear Unbiased Estimator) of β . Our derivation of equations { 5.4 } through { 5.7 } has actually established only that b^o is the (unrestricted) Normal-theory maximum likelihood estimator of β ; see Section §4a.2 of Rao(1973), pages 222-224, for the arguments that actually establish that all linear combinations, $P^T b^o$, coincide with the BLUE of the corresponding estimable linear combinations, $P^T \beta$.

5.2 The Likelihood of Mean Squared Error Optimality

Under the assumption that our “target values” for optimal shrinkage are given by the $\delta_i^{MSE} = \gamma_i^2/(\gamma_i^2 + \sigma^2 \lambda_i^{-1})$ factors of equation { 4.6 }, we now wish to explore methods for identifying numerical values of shrinkage-regression-factors that are most likely to be ON TARGET under Normal-theory. Rather than simply paraphrase the arguments given in my papers, Obenchain(1975,1981), I use a new approach here that I hope will be somewhat easier to follow. To minimize potential for misinterpretation, let me describe the general point-of-view assumed below:

- (i) Unrestricted maximum likelihood estimation of the parameters of a multiple regression model “uses up” a total of $R+2$ degrees-of-freedom. The estimate $\mathbf{b}^0 = \mathbf{X}^+ \mathbf{y} = \mathbf{G} \mathbf{c}$ (of the coefficient vector β) corresponds to R degrees-of-freedom. The estimates $\hat{\mu} = \bar{y} - \bar{x}^T \mathbf{b}^0$ (of the y intercept μ) and $\hat{\sigma}^2 = \mathbf{y}^T \mathbf{y} \cdot (1-R^2) / N$ (of the residual variance σ^2) use 1 degree-of-freedom each.
- (ii) Shrinkage regression estimation as viewed here always ends up “using” AT LEAST these same $R+2$ degrees-of-freedom. After all, shrinkage estimators are of the general form $\mathbf{b}^\star = \mathbf{G} \Delta \mathbf{c}$, where the \mathbf{c} vector (containing the least-squares estimates of the true uncorrelated components γ of β) again corresponds to R degrees-of-freedom. Similarly, $\mu^\star = \bar{y} - \bar{x}^T \mathbf{b}^\star$ uses up 1 degree-of-freedom. The residual sum-of-squares for shrinkage estimates, $(\mathbf{y} - \mathbf{X} \mathbf{b}^\star)^T (\mathbf{y} - \mathbf{X} \mathbf{b}^\star)$, can only exceed the minimum value, $\mathbf{y}^T \mathbf{y} \cdot (1-R^2)$, attained by unrestricted maximum likelihood. This excess lack-of-fit is of no real help in estimating the residual variance σ^2 , so the shrinkage regression estimate of σ^2 “usually” defaults back to the least-squares estimate, $\hat{\sigma}^2 = \mathbf{y}^T \mathbf{y} \cdot (1-R^2) / N$, with 1 degree-of-freedom.
- (iii) Maximum likelihood methods for identifying minimal mean-squared-error shrinkage may “use up” as few as only 1 or 2 more degrees-of-freedom than the minimum number, $R+2$, listed above. Specifically, the elements of the diagonal matrix, Δ , of shrinkage factors employed in $\mathbf{b}^\star = \mathbf{G} \Delta \mathbf{c}$ may be functions of only 1 or 2 parameters. (For example, in Section §5.3, these two parameters will be Q and k ; Q determines the shape/curvature of the shrinkage path, and k determines the extent of shrinkage along that path.) In any case, practical applications of shrinkage regression should always be thought of as using up a total of $R+3$, or $R+4$, or perhaps even more degrees-of-freedom. After all, they employ (restricted) estimates of Δ and μ as well as unrestricted estimates of the implied γ and σ^2 .
- (iv) The minimal mean-squared-error target value for shrinkage, Δ^{MSE} , is a nonlinear function of γ and σ^2 . As a result, maximum likelihood search over the HIGHLY RESTRICTED parameter space defining shrinkage factors may “use up” as few as only 3 or 4 degrees-of-freedom. Three or four degrees-of-freedom is frequently much less than the MINIMUM of $R+2$ employed in an unrestricted maximum likelihood search. The estimates of γ and σ^2 , denoted here by γ^{**} and σ^{**2} , that are derived within this restricted search are usually of very little interest in their own right. Instead, they are mere precursors that help us identify not only the restricted Δ factor values most likely to be Δ^{MSE} but also the minimum value for the corresponding minus-twice-log-likelihood-ratio (restricted likelihood divided by unrestricted likelihood.)

The Normal-theory likelihood for the uncorrelated components vector $\gamma = \mathbf{G}^T \beta$ and the residual variance σ^2 [evaluated, again, at $\mu = \bar{y} - \bar{x}^T \mathbf{G} \gamma$] is

$$L(\gamma, \sigma^2) = (2\pi\sigma^2)^{-N/2} e^{-\mathbf{u}^2/2\sigma^2}, \quad \{ 5.9 \}$$

where u^2 is the quadratic form

$$u^2 = (y - H \Lambda^{1/2} \gamma)^T (y - H \Lambda^{1/2} \gamma). \quad \{ 5.10 \}$$

Now note that $\delta_i^{\text{MSE}} = \gamma_i^2 / (\gamma_i^2 + \sigma^2 \lambda_i^{-1})$ can be rewritten as

$$\gamma_i^2 \cdot \lambda_i = \sigma^2 \cdot [\delta_i^{\text{MSE}} / (1 - \delta_i^{\text{MSE}})] \quad \{ 5.11 \}$$

or, equivalently, as

$$\gamma_i = \pm 1 \cdot \sigma \cdot \sqrt{\delta_i^{\text{MSE}} / [\lambda_i \cdot (1 - \delta_i^{\text{MSE}})]}. \quad \{ 5.12 \}$$

In other words, knowing the numerical values of $\delta_1^{\text{MSE}}, \delta_2^{\text{MSE}}, \dots, \delta_R^{\text{MSE}}$ would be tantamount to knowing the RELATIVE MAGNITUDES of the ABSOLUTE VALUES of the uncorrelated components, $|h_1|, |h_2|, \dots, |h_R|$, of β .

Equations { 5.11 } and/or { 5.12 } suggest the following FUNDAMENTAL DEFINITION for the likelihood that any given set of numerical values for shrinkage factors, $\delta_1, \delta_2, \dots, \delta_R$, coincide with the optimal mean-squared-error target values $\delta_1^{\text{MSE}}, \delta_2^{\text{MSE}}, \dots, \delta_R^{\text{MSE}}$. This likelihood is defined to equal the likelihood that

$$\gamma_i^{**} = \pm 1 \cdot \sigma^{**} \cdot \sqrt{\delta_i / [\lambda_i \cdot (1 - \delta_i)]}, \quad \{ 5.13 \}$$

where this likelihood has been maximized by choice of the R numerical signs (positive or negative) of the corresponding uncorrelated component estimates and by choice of estimate, σ^{**} , for the residual standard deviation. As explained above, these γ_i^{**} and σ^{**2} estimates usually are of little interest themselves, at least when the δ_i factors have been restricted to lie within a 1- or 2-parameter shrinkage family. But { 5.13 } is the general expression that would apply even if the δ_i were totally unrestricted; see subsection §5.2.1 below for a description of the "cubic" estimator that results in this totally unrestricted case.

Note that relationship { 5.13 } allows us to rewrite the quadratic form, u^2 of equation { 5.10 }, as

$$u^2 = (y - H S \xi \sigma)^T (y - H S \xi \sigma), \quad \{ 5.14 \}$$

where S is a diagonal matrix of signs (i.e. diagonal elements of ± 1) and ξ is the vector of values defined by the following positive square-roots:

$$\xi_i = \sqrt{\delta_i / (1 - \delta_i)}, \quad \{ 5.15 \}$$

for $1 \leq i \leq R$. Note, specifically, that the resulting maximum likelihood estimator of γ is being restricted to be of the general form

$$\gamma^{**} = \sigma^{**} \cdot S \Lambda^{-1/2} \xi \quad \{ 5.16 \}$$

whenever the shrinkage factors that define ξ yield minimum mean-squared-error. Before continuing, let us note that relationships { 5.15 } and { 5.16 } are well defined only when all δ_i factors are strictly less than 1; after all, $\delta_i = 1$ would imply $\xi_i = +\infty$. In other words, the possibility “no-shrinkage-at-all” along any principal axis is automatically being excluded from consideration as a potential mean-squared-error-optimal extent for shrinkage. On the other hand, values that are (numerically) very close to 1 are not being rejected out of hand. Thus, whenever a relatively large shrinkage factor value like $\delta_i = 0.95$ or 0.99 is found to be “most likely” in the sense defined below, there may be no real difference of any practical (numerical) importance between the corresponding least-squares and optimally-shrunk estimators.

Noting that $H^T H = I$, $S^2 = I$, and $y^T H = \sqrt{y^T y} \cdot r^T$ as in { 2.15 } and { 2.16 }, we now rewrite u^2 of { 5.10 } and { 5.14 } again as

$$u^2 = y^T y - 2 \cdot \sqrt{y^T y} \cdot r^T S \xi \sigma^{**} + \sigma^{**2} \xi^T \xi. \quad \{ 5.17 \}$$

To minimize u^2 , the R numerical signs in S should be chosen to make the middle, negative term as large as possible...assuming, of course, that the σ^{**} estimate is strictly positive. Clearly, the $r^T S \xi = \sum r_{yi} \cdot s_i \cdot \sqrt{\delta_i / (1 - \delta_i)}$ factor is maximized by choice of these signs when $s_i = \text{sign}(r_{yi})$, yielding $r^T S \xi = \sum |r_{yi}| \cdot \sqrt{\delta_i / (1 - \delta_i)}$.

Finding the optimal, strictly positive estimate, σ^{**} , of the residual standard deviation is the final, remaining step in minimizing the restricted minus-twice-log-likelihood. The corresponding partial derivatives are

$$\partial[-2 \cdot \ln(L^{**})] / \partial \sigma^{**} = \frac{2 \cdot N}{\sigma^{**}} - \frac{2 \cdot y^T y}{\sigma^{**3}} + \frac{2 \cdot \sqrt{y^T y} \cdot \sum |r_{yi}| \cdot \xi_i}{\sigma^{**2}}, \quad \{ 5.18 \}$$

and

$$\partial^2[-2 \cdot \ln(L^{**})] / \partial \sigma^{**2} = -\frac{2 \cdot N}{\sigma^{**2}} + \frac{6 \cdot y^T y}{\sigma^{**4}} - \frac{4 \cdot \sqrt{y^T y} \cdot \sum |r_{yi}| \cdot \xi_i}{\sigma^{**3}}. \quad \{ 5.19 \}$$

Equating the first derivative to zero yields the three solutions $\sigma^{**} = +\infty$ and

$$\sigma^{**} = \sqrt{y^T y} \cdot \frac{-\sum |r_{yi}| \cdot \xi_i \pm \sqrt{(\sum |r_{yi}| \cdot \xi_i)^2 + 4 \cdot N}}{2 \cdot N}. \quad \{ 5.20 \}$$

The negative solution is of no interest; this choice would make the middle term of { 5.17 } positive. And the infinite solution corresponds to an indeterminate second derivative of zero. The second derivative is strictly positive at the positive solution of { 5.20 } and, thus, minimum minus-twice-log-likelihood is achieved there, yielding:

$$\sigma^{**} = \frac{\sqrt{y^T y}}{2 \cdot N} \cdot \left[\sqrt{(\sum |r_{yi}| \cdot \xi_i)^2 + 4 \cdot N} - \sum |r_{yi}| \cdot \xi_i \right]$$

$$= \frac{2 \cdot \sqrt{y^T y}}{\left[\sqrt{(\sum |r_{yi}| \cdot \xi_i)^2 + 4 \cdot N + \sum |r_{yi}| \cdot \xi_i} \right]} \quad \{ 5.21 \}$$

Of course, the resulting restricted minimum of the likelihood in { 5.9 } cannot be smaller than the unrestricted minimum, { 5.4 } . Therefore, the resulting (non-negative) minus-twice-log-likelihood-ratio statistic will be of the form

$$- 2 \cdot \ln(L^{**} / \hat{L}) = N \cdot \ln(\sigma^{**2} / \hat{\sigma}^2) + \xi^T \xi - \sqrt{y^T y} \cdot \sum |r_{yi}| \cdot \xi_i / \sigma^{**}, \quad \{ 5.22 \}$$

and this statistic will have an asymptotic chi-squared distribution (as N increases to ∞) with degrees-of-freedom equal to R minus the number of “free” parameters remaining among the $\delta_1, \delta_2, \dots, \delta_R$ factors under any restriction that might be imposed.

5.2.1 Unrestricted Maximum Likelihood Shrinkage: The Cubic Estimator

When no restrictions whatsoever are placed upon the δ -factors, { 5.22 } will be zero and will have zero degrees-of-freedom. To see this, note that the unrestricted maximum likelihood estimate of $\delta_i^{\text{MSE}} = \gamma_i^2 / (\gamma_i^2 + \sigma^2 \lambda_i^{-1})$ is clearly $c_i^2 / (c_i^2 + \hat{\sigma}^2 \lambda_i^{-1}) = r_{yi}^2 / [r_{yi}^2 + (1 - R^2)/N]$ where $c = G^T b^0$ is the vector of least-squares estimates for uncorrelated components and $\hat{\sigma}^2 = y^T y \cdot (1 - R^2)/N$ is the least-squares residual mean square of { 5.4 } . The corresponding values of the ξ_i terms are $\xi_i = |r_{yi}| \cdot \sqrt{N/(1 - R^2)}$. Therefore

$$\sum |r_{yi}| \cdot \xi_i = R^2 \cdot \sqrt{N/(1 - R^2)} \quad \text{and} \quad \sqrt{(\sum |r_{yi}| \cdot \xi_i)^2 + 4 \cdot N} = (2 - R^2) \cdot \sqrt{N/(1 - R^2)} .$$

In other words, the unrestricted estimates are $\sigma^{**2} = \hat{\sigma}^2$ in { 5.21 } and $\gamma^{**} = c$ in { 5.16 } .

The corresponding maximum-likelihood shrinkage estimator, $\hat{\delta}_i c_i$, for γ_i would be $c_i^3 / (c_i^2 + \hat{\sigma}^2 \lambda_i^{-1})$, which is a specific nonlinear estimator of “cubic” form. Thompson(1968), Figures 1 and 2 [pages 116 and 117], gave Normal-theory mean-squared-error plots (computed using numerical integration) for this special form of nonlinear estimator, where his horizontal axis was $|\gamma|/\sigma$ and his plotting range was unbounded (0 to ∞ .) Dwivedi, Srivastava, and Hall(1980) and Hemmerle and Carey(1981) also study the mean-squared-error properties of this cubic estimator. See Figure XX, Chapter 6, for a plot of the simulated mean-squared-error-risk of the cubic estimator versus δ^{MSE} , i.e. over the finite range from 0 to 1.

5.2.2 Maximum Likelihood UNIFORM Shrinkage

Under the “uniform shrinkage” restriction that $\delta_1 = \delta_2 = \dots = \delta_R$, this common shrinkage factor can be written as $\delta = 1 / (1 + k)$. As a result, $\xi_1 = \xi_2 = \dots = \xi_R = \sqrt{\delta / (1 - \delta)} =$

$k^{-1/2}$, and the value of k that minimizes { 5.22 } [with chi-square degrees-of-freedom = $R - 1$ in the limit as n approaches ∞] is

$$k^{**} = (1 - R \cdot \overline{|r|^2}) / (N \cdot \overline{|r|^2}), \quad \{ 5.23 \}$$

where $\overline{|r|} = (|r_{y1}| + |r_{y2}| + \dots + |r_{yR}|) / R$ is the average of the absolute values of the principal correlations, Obenchain(1975). The corresponding minimum minus-twice-log-likelihood-ratio is then $-2 \cdot \ln(L^{**} / \hat{L}) = N \cdot \ln[1 + \frac{(R-1)}{(N-R-1)} S]$, where S is the "uniform shrinkage statistic" of Obenchain(1975) :

$$S = \frac{(N-R-1) \cdot \sum (|r_{yi}| - \overline{|r|})^2}{(R-1) \cdot (1 - R^2)}. \quad \{ 5.24 \}$$

No derivation of equations { 5.23 } and { 5.24 } will be given now because they are simple special cases [$Q=1$] of the general theory derived below in Section §5.3.

The corresponding restricted estimates of σ and of the uncorrelated components of β would be

$$\sigma^{**2} = y^T y \cdot (1 - R \cdot \overline{|r|^2}) / N \quad \{ 5.25 \}$$

and

$$\gamma_i^{**} = \sqrt{y^T y / \lambda_i} \cdot \text{sign}(r_{yi}) \cdot \overline{|r|} \quad \dots \text{for } 1 \leq i \leq R. \quad \{ 5.26 \}$$

Again, these latter restricted estimates (the variance and the components) are of little real interest. After all, the numerical value of the maximum-likelihood "common" shrinkage factor, $\delta^{**} = N \cdot \overline{|r|^2} / [1 + (N - R) \cdot \overline{|r|^2}]$, would actually be used in conjunction with the UNRESTRICTED maximum-likelihood component estimates (c_1, c_2, \dots, c_R) and the UNRESTRICTED residual variance, $\hat{\sigma}^2$, of { 5.4 }. In other words, the restricted σ^{**2} and γ_i^{**} estimates are of interest ONLY in the sense that they help to define the minimum minus-twice-log-likelihood-ratio via equation { 5.24 }.

At the time of my original publication on Normal-theory maximum likelihood methods for shrinkage regression, Obenchain(1975), the only known closed-form solutions to the general expressions { 5.16 } and { 5.21 } were the two special cases treated above in subsections §5.2.1 and §5.2.2. Thus, I proposed a general technique I called likelihood monitoring for applying equations { 5.16 } and { 5.21 } to any parametric family of generalized shrinkage factors. This approach simply involves actual numerical computation of { 5.16 }, { 5.21 } and { 5.22 } upon a lattice of numerical values for $\delta_1, \delta_2, \dots, \delta_R$. These sorts of computations can be tedious, of course, but the vast majority of criteria that have been proposed for choosing a "best" shrinkage estimator are commonly applied in this computationally-intensive fashion.

5.3 Closed Form Expressions within the 2-Parameter Family

Within the 2-parameter family of Goldstein and Smith(1974), shrinkage factors are of the general form $\delta_i = 1 / (1 + k \cdot \lambda_i^{Q-1})$ of equation { 3.8 }, where the power, Q, determines the SHAPE/CURVATURE of the shrinkage path through likelihood space while the k factor determines the EXTENT of shrinkage. The natural range for the k parameter is all the way from k=0 for “no shrinkage” to the limit as k approaches $+\infty$, where all δ_i factors are shrunken “completely” to 0. The special case of shape Q=1 for “uniform shrinkage” was described in subsection §5.2.2. Among the other common choices for Q described in Section §3.3 of Chapter 3 are Q=0 for “ordinary ridge regression,” Hoerl and Kennard(1970), and the limit as Q approaches $-\infty$ for “principal components regression,” Marquardt(1970). In practical applications to ill-conditioned regression problems where the eigenvalue spectrum of the regressor $X^T X$ matrix is wide, values of Q outside of the range $-5 \leq Q \leq +5$ rarely need to be considered. At the other extreme, where absolutely no ill-conditioning is present because $\lambda_1 = \lambda_2 = \dots = \lambda_R$, all values of Q would yield the same, uniform shrinkage pattern.

5.3.1 The most-likely-to-be-mse-optimal shrinkage extent, k, for given shape/curvature.

When $\delta_i = 1 / (1 + k \cdot \lambda_i^{Q-1})$, the ξ_i terms of { 5.15 } are of the general form $\xi_i = \sqrt{\delta_i / (1 - \delta_i)} = \sqrt{\lambda_i^{(1-Q)} / k}$. Therefore { 5.16 } becomes

$$\gamma_i^{**} = \pm \sigma^{**} / \sqrt{k \cdot \lambda_i^{Q/2}}. \quad \{ 5.27 \}$$

There is a redundancy in { 5.27 } between k and the estimate of σ that could not be fully exploited in deriving a general expression like that given by equation { 5.21 }.

Let us now denote the common, unknown value of $\gamma_i^2 \lambda_i^Q = \sigma^{**2} / k$ in { 5.27 } by ρ_Q^2 :

$$\rho_Q^2 = \gamma_1^2 \lambda_1^Q = \gamma_2^2 \lambda_2^Q = \dots = \gamma_R^2 \lambda_R^Q. \quad \{ 5.28 \}$$

Equation { 5.27 } can then be rewritten as $\gamma_i^{**} = \pm \rho_Q^{**} \cdot \lambda_i^{-Q/2}$ and, again using $s_i = \text{sign}(r_{yi})$, the general residual-sum-of-squares equation of { 5.10 }, { 5.14 } and { 5.17 } now becomes

$$u^2 = y^T y - 2 \cdot \sqrt{y^T y} \cdot \rho_Q^{**} \cdot \sum |r_{yi}| \cdot \lambda_i^{(1-Q)/2} + \rho_Q^{**2} \cdot \sum \lambda_i^{(1-Q)}. \quad \{ 5.29 \}$$

The corresponding partial derivatives of u^2 are

$$\partial[u^2] / \partial \rho_Q^{**} = -2 \cdot \sqrt{y^T y} \cdot \sum |r_{yi}| \cdot \lambda_i^{(1-Q)/2} + 2 \cdot \rho_Q^{**} \cdot \sum \lambda_i^{(1-Q)}, \quad \{ 5.30 \}$$

and

$$\partial^2[u^2] / \partial \rho_Q^{**2} = +2 \cdot \sum \lambda_i^{(1-Q)}. \quad \{ 5.31 \}$$

Thus the minimum value of u^2 clearly occurs at $\partial[u^2]/\partial\rho_Q^{**} = 0$, which implies

$$\rho_Q^{**} = \sqrt{y^T y} \cdot \frac{\sum |r_{yi}| \cdot \lambda_i^{(1-Q)/2}}{\sum \lambda_i^{(1-Q)}} . \quad \{ 5.32 \}$$

and

$$u^{**2} = \text{minimum}(u^2) = y^T y \cdot [1 - R^2 \cdot \text{CRL}^2(Q)] , \quad \{ 5.33 \}$$

where $\text{CRL}(Q)$ is the ‘‘curlicue’’ function that measures CORRELATION between the vector of absolute values of the principal correlations and the vector of regressor singular values raised to the $(1 - Q)/2$ -th power. Specifically,

$$\text{CRL}(Q) = \frac{\sum |r_{yi}| \cdot \lambda_i^{(1-Q)/2}}{\sqrt{\sum r_{yi}^2 \cdot \sum \lambda_i^{(1-Q)}}} . \quad \{ 5.34 \}$$

Note that this correlation can also be viewed as the Cosine of the angle between the ‘‘R-vector’’ of absolute principal axis correlations and the ‘‘L-vector’’ of regressor eigenvalues raised to a power determined by the path shape/curvature parameter, Q ; this notation motivates the $\text{CRL}(Q)$ mnemonic, Obenchain(1981). [In Obenchain(1975), equation (4.6), $\text{CRL}(Q)$ was denoted by $\text{COS}(q)$.]

Our final step in maximizing the restricted Normal-theory likelihood is to choose the estimator of σ^{**2} given the minimum u^2 , which is the exact same sort of problem we treated in equations { 5.7 } and { 5.8 }. Of the two possible solutions to $\partial[-2 \cdot \ln(L^{**})]/\partial(\sigma^{**2}) = 0$, the $\sigma^{**2} = +\infty$ solution is again ruled out in favor of $\sigma^{**2} = u^{**2}/N$. As a result, the most-likely-to-be-optimal extent of shrinkage along the path of shape Q is given by:

$$k^{**} = \sigma^{**2}/\rho_Q^{**2} = \left[\sum \lambda_i^{(1-Q)} \right] \cdot \frac{[1 - R^2 \cdot \text{CRL}^2(Q)]}{[N \cdot R^2 \cdot \text{CRL}^2(Q)]} . \quad \{ 5.35 \}$$

The corresponding minimum minus-twice-log-likelihood-ratio is then $-2 \cdot \ln(L^{**}/\hat{L}) = N \cdot \ln[1 + \frac{(R-1)}{(N-R-1)} S(Q)]$, where $S(Q)$ becomes :

$$S(Q) = \frac{(N-R-1) \cdot R^2 [1 - \text{CRL}^2(Q)]}{(R-1) \cdot (1 - R^2)} . \quad \{ 5.36 \}$$

Note that { 5.36 } agrees with equation (4.6) of Obenchain(1975), but the closed-form expression, { 5.35 }, was unknown at that time. This most-likely-to-be-mse-optimal value of k given Q was first derived in Obenchain(1981), equation (2.5).

5.3.2 The most-likely-to-be-mse-optimal shrinkage shape/curvature, Q .

It is clear from { 5.36 } that the minimum minus-twice-log-likelihood-ratio depends upon Q only through the curlicue function of { 5.34 } :

$$\text{CRL}(Q) = \frac{\sum |r_{yi}| \cdot \lambda_i^{(1-Q)/2}}{\sqrt{\sum r_{yi}^2 \cdot \sum \lambda_i^{(1-Q)}}} .$$

And it is clear from this definition that CRL(Q) cannot be made negative by choice of Q. As a result, S(Q) is minimized (as is u^{**2} of { 5.33 }) by choice of Q by making CRL(Q) as large as possible.

Before going further, perhaps we should discuss why choosing the Q-shape so as to maximize the curlicue function represents intuitive "common sense." Notice, first, that the ordered regressor eigenvalues, $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_R > 0$, will have very little influence upon CRL(Q) whenever Q is close to +1 ...because the $\lambda_i^{(1-Q)/2}$ values are all equal to 1 at Q=+1. Whenever the absolute values of all of the principal correlations are "nearly" equal, the angle between the |R|-vector and the Q=+1 L-vector $\propto 1$ will be small and, thus, CRL(Q) will be maximized at a Q value close to +1. And Q=+1 represents uniform shrinkage, $\delta_1 = \delta_2 = \dots = \delta_R$.

Next, note that the shrinkage factors will be monotonically non-increasing, $\delta_1 \geq \delta_2 \geq \dots \geq \delta_R$, when Q is strictly less than +1 and monotonically non-decreasing, $\delta_1 \leq \delta_2 \leq \dots \leq \delta_R$, when Q is strictly greater than +1. After all, these factors are being restricted to be of the functional form $\delta_i = 1 / [1 + k \cdot \lambda_i^{(1-Q)}]$.

Therefore, when the "trailing" absolute principal correlations [$|r_{yR}|, |r_{y(R-1)}|, \dots$] are relatively large, CRL(Q) will tend to be maximized at values of Q greater than +1. These are the somewhat pathological cases where the regressor coordinates that have the least adequate spread in their numerical values are the coordinates most highly correlated with the response.

But, when the "leading" absolute principal correlations [$|r_{y1}|, |r_{y2}|, \dots$] are relatively large, CRL(Q) will tend to be maximized at values of Q less than +1. In fact, the Q that maximizes CRL(Q) may be less than 0 in these cases. These are the "business-as-usual" cases where the regressor coordinates that have the most adequate spread in their numerical values are most highly correlated with the response. After all, this is the strategy you would use in a "designed experiment" ...where you would deliberately explore responses over a relatively wide range for any/all relatively important "factors."

In actual applications of normal-theory maximum-likelihood to ill-conditioned regression problems, I favor considering only a limited number of possible shapes, Q. For example, my personal computer application, RXridge, considers only integer and half-integer values within the range $-5 \leq Q \leq +5$. I see no practical reason for ever considering a finer lattice of Q shapes than, say, 0.1 (one place after the decimal) or a wider range of Q shapes than $-5 \leq Q \leq +5$; but this is, perhaps, mostly a matter of personal taste. Anyway, whenever a

limited number of Q shapes are under consideration, the most straight-forward way to find the corresponding restricted maximum of CRL(Q) is simply to compute all of these values ...then pick the Q shape yielding the largest CRL(Q) value.

For those shrinkage regression practitioners who simply cannot resist the temptation to locate "the" optimal Q shape (with great numerical precision), a Newtonian descent method for iterative search can be used. Specifically, with Q_s^{**} denoting the best estimate of the Q that maximizes CRL(Q) at stage s of the iteration, the update equation for step s+1 becomes

$$Q_{s+1}^{**} = Q_s^{**} - [CRL'(Q) / CRL''(Q)], \quad \{ 5.37 \}$$

where

$$CRL'(Q) = \partial CRL(Q) / \partial Q = \sum_{i=1}^R \left[\frac{|r_{yi}| \cdot \lambda_i^{(1-Q)/2}}{R \cdot \sqrt{\sum \lambda_i^{(1-Q)}}} \cdot \left(H_i + \frac{J}{2} \right) \right], \quad \{ 5.38 \}$$

for

$$R^2 = \sum r_{yj}^2, \quad H_i = \frac{-\ln(\lambda_i)}{2}, \quad J = \frac{\sum \lambda_j^{(1-Q)} \cdot \ln(\lambda_j)}{\sum \lambda_j^{(1-Q)}},$$

and

$$CRL''(Q) = \partial^2 CRL(Q) / \partial Q^2 = \sum_{i=1}^R \left[\frac{|r_{yi}| \cdot \lambda_i^{(1-Q)/2}}{R \cdot \sqrt{\sum \lambda_i^{(1-Q)}}} \cdot \left\{ \left(H_i + \frac{J}{2} \right)^2 + \frac{J^2}{2} - K \right\} \right], \quad \{ 5.39 \}$$

for

$$K = \frac{\sum \lambda_j^{(1-Q)} \cdot \ln^2(\lambda_j)}{\sum \lambda_j^{(1-Q)}}.$$

Convergence to a (possibly local) maximum of CRL(Q) requires finding a shape value Q^{**} such that $CRL'(Q^{**})=0$ and $CRL''(Q^{**}) < 0$. The step-size, $[CRL'(Q)/CRL''(Q)]$, in { 5.37 } should be bisected whenever $CRL(Q_{s+1}^{**})$ fails to achieve an increase over $CRL(Q_s^{**})$, and the search direction should be reversed when $CRL''(Q_s^{**}) > 0$.

5.3.3 The limit as the shrinkage shape/curvature, Q, approaches $-\infty$.

Note that CRL(Q) can also be thought of as the cosine of the angle between the vector of absolute principal correlations, $[|r_{y1}|, |r_{y2}|, \dots, |r_{yR}|]$, and the following vector of powers of eigenvalue ratios, $[1, (\lambda_2/\lambda_1)^{(1-Q)/2}, (\lambda_3/\lambda_1)^{(1-Q)/2}, \dots, (\lambda_R/\lambda_1)^{(1-Q)/2}]$. This latter vector clearly approaches $[1, 0, 0, \dots, 0]$ as Q approaches $-\infty$ whenever $\lambda_1 > \lambda_2$, i.e. when the leading eigenvalue of the centered-regressor $X^T X$ matrix is larger than all of the other eigenvalues. Thus CRL(Q) approaches $|r_{y1}|/R$ as Q approaches $-\infty$; similarly, $k^{**} \cdot \lambda_1^{(Q-1)}$

approaches $[1 - r_{y1}^2] / [N \cdot r_{y1}^2]$ while $k^{**} \cdot \lambda_j^{(Q-1)}$ approaches $+\infty$ for $j = 2, 3, \dots, R$. The shrinkage factors most-likely-to-be-mse-optimal in this limit are thus $\delta_1^{**}(-\infty) = N / [N - 1 + r_{y1}^2]$ and $\delta_2^{**}(-\infty) = \delta_3^{**}(-\infty) = \dots = \delta_R^{**}(-\infty) = 0$, which is a point on the principal components regression path, Massy(1965), that has a Marquardt(1970) fractional rank of less than 1.

5.3.4 Large Sample Chi-Squared Tests of MSE-Optimality

A large sample χ^2 (Chi-Squared) test can be based upon the minimum value of the minus-twice-log-likelihood-ratio, $-2 \cdot \ln(L^{**} / \hat{L}) = N \cdot \ln[1 + \frac{(R-1)}{(N-R-1)} S(Q)]$, where $S(Q)$ is defined as in { 5.36 }. The degrees-of-freedom used in this test would be $(R - 1)$ if one's shrinkage shape parameter, Q , had been selected without reference to the observed response data. But the appropriate degrees-of-freedom would be $(R - 2)$ if, instead, $S(Q)$ has been minimized by choice of Q . Whenever this χ^2 statistic is significantly greater than zero, statistical evidence has been accumulated suggesting that the 2-parameter family is "too restrictive" to contain the MSE-optimal values for the shrinkage factors.

5.4 Maximum Likelihood Methods for Mixed Linear Models

Statistical literature on the subject of mixed linear models (i.e. models containing both fixed and random coefficients) has been growing for 40-50 years; several major, new contributions to this area have appeared within the last 25 years. Henderson(1950) introduced the mixed model equations; his more recent fundamental contributions, Henderson(1975, 1984, 1990), include BLUP theory. Rao(1971a,b) introduced MINQUE and MIVQUE estimates of variance components; Patterson and Thompson(1971) defined REML estimation; Searle(1971, 1979, 1988) unified the theory of mixed, linear models and variance components; Harville(1977, 1988, 1990) has provided maximum likelihood theory and algorithms as well as prediction methodology; and Robinson(1990) has provided a highly readable BLUP review article. The vast majority of technical details on normal-distribution-theory maximum-likelihood estimation for mixed linear models will be postponed until Chapter 7 rather than being presented here in Chapter 5. However, we will introduce sufficient material, here in Chapter 5, to establish a few key parallels between the otherwise distinct maximum-likelihood approaches to fixed coefficient and random coefficient models.

A mixed linear model can be written in the general form:

$$y = X \cdot \beta + Z \cdot \theta + \eta \quad \{ 5.40 \}$$

where

$$\text{Var}(y) = V = Z \cdot \text{Var}(\theta) \cdot Z^T + \text{Var}(\eta) \quad \{ 5.41 \}$$

and the variance matrices, $G = \text{Var}(\theta)$ and $R = \text{Var}(\eta)$, are positive-definite matrices (frequently of block-diagonal form) that containing known or unknown parameters, generally called "variance components."

Now, the "unified" theory of mixed linear models tells us that the BLUE's and BLUP's are solutions to the Henderson(1975, 1984) MIXED MODEL EQUATIONS

$$\begin{bmatrix} \mathbf{X}^T \mathbf{R}^{-1} \mathbf{X} & \mathbf{X}^T \mathbf{R}^{-1} \mathbf{Z} \\ \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{X} & \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \cdot \begin{bmatrix} \hat{\beta} \\ \hat{\theta} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T \mathbf{R}^{-1} \mathbf{y} \\ \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{y} \end{bmatrix}. \quad \{ 5.42 \}$$

We can display closed form solutions to these equations under the assumption/pretense that the G and R matrices are known matrices. Namely, the BLUEs would be

$$\begin{aligned} \hat{\beta} &= (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{y}, & \{ 5.43 \} \\ &= [\mathbf{X}^T (\mathbf{R} + \mathbf{Z} \mathbf{G} \mathbf{Z}^T)^{-1} \mathbf{X}]^{-1} \mathbf{X}^T (\mathbf{R} + \mathbf{Z} \mathbf{G} \mathbf{Z}^T)^{-1} \mathbf{y}, \end{aligned}$$

and the BLUPs would be

$$\begin{aligned} \hat{\theta} &= \mathbf{G} \mathbf{Z}^T \mathbf{V}^{-1} [\mathbf{y} - \mathbf{X} \hat{\beta}], & \{ 5.44 \} \\ &= (\mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z} + \mathbf{G}^{-1})^{-1} [\mathbf{Z}^T \mathbf{R}^{-1} - \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{X} \mathbf{W} \mathbf{X}^T (\mathbf{R} + \mathbf{Z} \mathbf{G} \mathbf{Z}^T)^{-1}] \mathbf{y}, \end{aligned}$$

where $\mathbf{V} = \mathbf{R} + \mathbf{Z} \mathbf{G} \mathbf{Z}^T$ from { 5.41 } and $\mathbf{W} = \{ \mathbf{X}^T (\mathbf{R} + \mathbf{Z} \mathbf{G} \mathbf{Z}^T)^{-1} \mathbf{X} \}^{-1}$.

In his recent review article, Robinson(1991) stresses that Henderson's BLUP terminology is "reasonable" in the sense that $\hat{\theta}$ of { 5.44 } is indeed a Linear and Unbiased estimator of the random θ vector of { 5.6 } that is Best in a minimum variance sense...but only when the G and R matrices are formed using the UNKNOWN, TRUE values for all variance components! [By the way, $\hat{\theta}$ is termed a Predictor (rather than an estimator) primarily because θ is random rather than a fixed effect (unknown constant.)]

In reality, the variance components are frequently not only unknown but also the primary focus of one's attention in both estimation and statistical inference! Because numerical estimates for variance components (derived from the data at hand) are inserted into { 5.43 } and { 5.44 }, in practical applications neither $\hat{\beta}$ nor $\hat{\theta}$ is actually linear, neither $\hat{\beta}$ nor $\hat{\theta}$ is actually unbiased, and neither $\hat{\beta}$ nor $\hat{\theta}$ is actually best in any minimum variance sense! In other words, BLUE and BLUP terminology is truly "unfortunate" when applied to mixed model estimation.

An interesting facet of mixed model estimation that is implied by (but, perhaps, not immediately obvious from) equations { 5.42 }, { 5.43 } and { 5.44 } is that BLUPs are forms of “shrinkage estimators.” This analogy is, perhaps, most obvious in the following special case.

5.5 Completely Random Models with a Single Variance Component

Golub, Heath and Wahba(1979) display a (random coefficient) maximum likelihood criterion for picking the extent of shrinkage in ridge regression [their equation (5.3)] that I, at least, found quite mysterious until I studied Shumway(1982). The mixed model considered by these authors is “completely random” in the sense that the $X \cdot \beta$ term of { 5.40 } contains only the (rank 1) overall mean, $1 \cdot \mu$. Furthermore, the random θ coefficient variation involves a “single” (unknown) variance component, σ_θ^2 . Thus $\text{Var}(\theta) = G = \sigma_\theta^2 \cdot D^{-1}$ where D^{-1} represents the known dispersion structure of the unknown θ vector. In other words, all coefficients are assumed to have known intercorrelations and known relative variances. The special case of uncorrelated, homoscedastic coefficients is $D = I$.

In exactly the same way that the non-constant columns of X are usually “centered,” we suppose now that the columns of Z have been made to sum to 0 by subtracting off column means. To avoid complications unnecessary to this discussion, suppose that the centered Z matrix is of full (column) rank, P_z . Finally, suppose that the η disturbance terms are uncorrelated and homoscedastic: $R = \sigma^2 \cdot I$, as in { 2.2 }, where σ^2 is the unknown error variance component.

5.5.1 Demonstration that BLUP estimates are shrinkage estimates in this case.

Under the above assumptions, $X^T R^{-1} Z = 0$ in { 5.42 }, so that the matrix equations for $\hat{\mu}$ and $\hat{\theta}$ “uncouple” as follows. The top equation in { 5.42 } reduces to $\hat{\mu} = 1^T y / 1^T 1 = \bar{y}$, and the bottom P_z equations yield:

$$\hat{\theta} = (Z^T Z + (\sigma^2 / \sigma_\theta^2) \cdot D)^{-1} Z^T y. \quad \{ 5.45 \}$$

The second matrix expression in { 5.44 } is equivalent to { 5.45 } because $Z^T R^{-1} X = 0$. Notice also that, because the Z matrix has been centered, replacing the response vector y by $(y - 1 \cdot \bar{y})$ in { 5.45 } would not change the $\hat{\theta}$ estimate.

Now note in equation { 5.45 } that:

- (i) the variance component ratio, $\sigma^2 / \sigma_\theta^2 = \phi^{-2}$, plays the role of the “k” (shrinkage-extent) factor of { 3.9 } and { 5.27 }, while

(ii) D plays the role of the $(Z^T Z)^Q$ matrix in equation { 3.9 } for the “Q-shape” of equation { 5.27 }.

In particular, $D = I$ in { 5.45 } yields the random-coefficient version of the (ordinary) ridge regression shrinkage path [$Q=0$] of Hoerl and Kennard(1970). Therefore, we have established that BLUP estimates are shrinkage estimates ...at least in the case of random coefficient models with a single variance component. That more complicated forms of BLUP also represent shrinkage can be easily verified via numerical computation.

5.5.2 Random coefficient maximum likelihood choice of shrinkage extent.

The Normal-theory joint likelihood function for the responses, y , can be written in the form

$$L(\theta, \sigma_\theta^2, \sigma^2) = (2\pi|V|)^{-1/2} e^{-u^2/2}, \quad \{ 5.46 \}$$

where u^2 is the quadratic form

$$u^2 = (y - 1 \cdot \mu)^T V^{-1} (y - 1 \cdot \mu), \quad \{ 5.47 \}$$

and $V = \sigma_\theta^2 \cdot Z D^{-1} Z^T + \sigma^2 \cdot I$ from { 5.41 } .

Writing $k = \sigma^2 / \sigma_\theta^2$ and using well-known determinant and matrix-inverse identities [Rao(1973), pages 32 and 33], it follows that

$$|D| \cdot |V| = \begin{vmatrix} \sigma^2 I & \sigma_\theta Z \\ -\sigma_\theta Z^T & D \end{vmatrix} = \sigma^{2 \cdot N} \cdot |D + k^{-1} \cdot Z^T Z|, \quad \{ 5.48 \}$$

and

$$V^{-1} = \sigma^{-2} \cdot I - \sigma^{-2} \cdot Z (Z^T Z + k \cdot D)^{-1} Z^T. \quad \{ 5.49 \}$$

These expressions allow us to rewrite { 5.46 } as

$$-2 \cdot \ln(L) = \ln(2\pi) + N \cdot \ln \sigma^2 - \ln |D| - P_z \cdot \ln k + \ln |Z^T Z + kD| + u^2 / \sigma^2, \quad \{ 5.50 \}$$

where $u^2 = \sum (y_j - \mu)^2 - y^T Z (Z^T Z + k \cdot D)^{-1} Z^T y$. This minus-twice-log-likelihood is minimized, first, by taking $\hat{\mu} = \bar{y}$ to minimize u^2 for any given value of k . Then, exactly as in equations { 5.7 } and { 5.8 }, the minimizing error variance component is $\hat{\sigma}^2 = \hat{\sigma}^2(k) = [\sum (y_j - \bar{y})^2 - y^T Z \hat{\theta}(k)] / N$ for $\hat{\theta}(k) = \hat{\theta}$ of { 5.45 }. Substituting these values into { 5.50 } yields an expression for the minus-twice-log-likelihood that is a function of k only :

$$-2 \cdot \ln(L) = \ln(2\pi) + N - \ln |D| + N \cdot \ln \hat{\sigma}^2(k) - P_z \cdot \ln k + \ln |Z^T Z + k \cdot D|, \quad \{ 5.51 \}$$

as in equation (18) of Shumway(1982). Numerical search over a lattice of alternative values for k would then be used to locate the (approximate) minimum of { 5.51 }.

Rather than base computations on this minus-two-log-likelihood expression, Golub, Heath and Wahba(1979) suggest minimizing the equivalent criterion:

$$M(k) = \frac{1}{N} \cdot \frac{(y - \bar{y} \cdot 1)^T (I - A(k)) (y - \bar{y} \cdot 1)}{|I - A(k)|^{1/N}}, \quad \{ 5.52 \}$$

where $A(k) = Z (Z^T Z + k \cdot D)^{-1} Z^T$. Note that the numerator of { 5.52 } is simply $N \cdot \hat{\sigma}^2(k)$ and that $A(k) = I - \sigma^2 V^{-1}$ by { 5.49 }. Thus the N-th root of the determinant can be rewritten as $| I - A(k) |^{1/N} = |\sigma^2 V^{-1}|^{1/N}$ and this, in turn, is equivalent, by { 5.48 }, to the product of terms $[|D|^{1/N} \cdot k^{P_Z/N} / |D \cdot k + Z^T Z|^{1/N}]$ whose negative logarithm is contained in { 5.51 } when that expression is divided by N.

REFERENCES for Chapter Five

- Golub, G.H., Heath, M., and Wahba, G. (1979). "Generalized cross-validation as a method for choosing a good ridge parameter." **Technometrics** 21, 215-223.
- Henderson, C. R. (1950). "Estimation of genetic parameters (abstract.)" **Annals of Mathematical Statistics** 21, 309-310.
- Henderson, C. R. (1973). "Sire evaluation and genetic trends." In **Proceedings of the Animal Breeding and Genetics Symposium in Honor of Dr. Jay L. Lush** 10-41. Amer. Soc. Animal Sci. – Amer. Dairy Sci. Assoc. – Poultry Sci. Assn., Champaign, Illinois.
- Henderson, C. R. (1984). **Applications of Linear Models in Animal Breeding**, University of Guelph.
- Henderson, C. R. (1990). "Statistical methods in animal improvement: historical overview." In **Advances in Statistical Methods for Genetic Improvement in Livestock**. Springer-Verlag 1-14, 413-436.
- Obenchain, R. L. (1975). "Ridge analysis following a preliminary test of the shrunken hypothesis." **Technometrics**, 17, 431-441. (Discussion: McDonald, G. C., 443-445.)
- Obenchain, R. L. (1981). "Maximum likelihood ridge regression and the shrinkage pattern hypotheses." Abstract 81t-23. **I.M.S. Bulletin** 10, 37.
- Obenchain, R. L. (1984). "Maximum likelihood ridge displays." **Communications in Statistics A**, 13, 227-240. (Proceedings of the Fordham Ridge Symposium, ed. H. D. Vinod.)
- Rao, C. R. (1973). **Linear Statistical Inference and its Applications**, 2nd edition. New York: John Wiley & Sons.
- Searle, S. R. (1971). **Linear Models**. New York: John Wiley and Sons.
- Shumway, R. H. (1982). "Maximum likelihood estimation of the ridge parameter in linear regression." **Technical Report, Department of Statistics**, University of California at Davis.

Further Reading for Chapter Five

Dwivedi, T. D., Srivastava, V. K. and Hall, R. L. (1980). "Finite sample properties of ridge estimators." **Technometrics** 22, 205-212.

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

Fuller, W. A. and Battese, G. E. (1973). "Transformations for estimation of linear models with nested error structure," **Journal of the American Statistical Association**, 68, 626-632.

Harville, D. A. (1977). "Maximum likelihood approaches to variance component estimation and to related problems," **Journal of the American Statistical Association** 72, 320-338.

Harville, D. A. (1986). "Using least squares software to compute combined intra-interblock estimates of treatment contrasts," **The American Statistician**, 40, 153-157.

Hemmerle, W. J. and Carey, M. B. (1981). "Some properties of generalized ridge estimators." Department of Computer Science and Experimental Statistics, University of Rhode Island.

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

Jennrich, R. I. and Schluchter, M. D. (1986). "Unbalanced repeated-measures models with structured covariance matrices," **Biometrics**, 42, 805-820.

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

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

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

Robinson, G. K. (1991). "That BLUP is a good thing: the estimation of random effects," (with discussion.) **Statistical Science**, 6, 15-51.

Schluster, M. D. (1988). "Unbalanced repeated measures models with structured covariance matrices," **BMDP Statistical Software Manual**, vol.2, 1081-1114. University of California Press, Berkeley.

Searle, S. R. (1979). "Notes on variance component estimation: a detailed account of maximum likelihood and kindred methodology," Biometrics Unit Paper **BU-673-M**, Cornell University (149 pages.)

Searle, S. R. (1988). "Mixed models and unbalanced data: wherefrom, whereat, and whereto?" **Communications in Statistics - Theory and Methods**, 17, 935-968.

Thompson, J. R. (1968). "Some shrinkage techniques for estimating the mean." **Journal American Statistical Association** 63, 113-122.