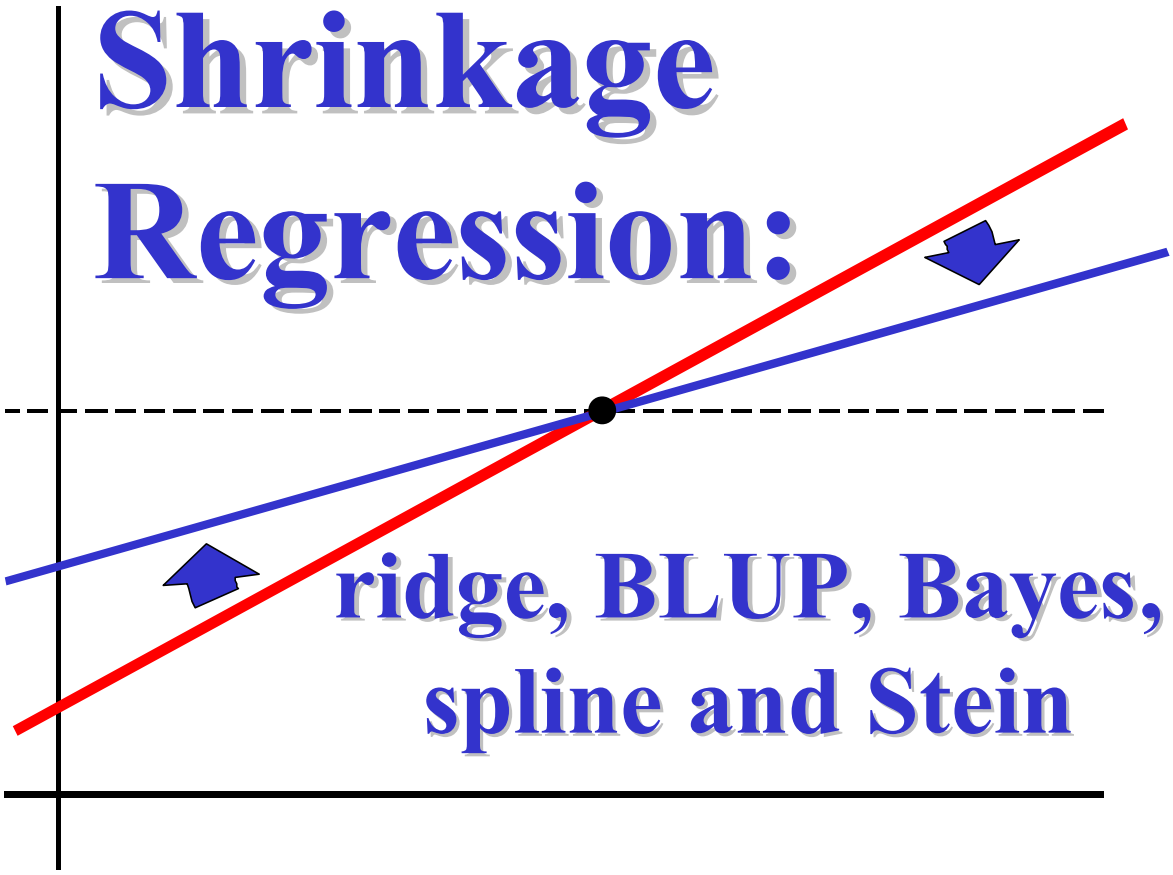


Chapter 08: Bayesian Formulations

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

Chapter 8: BAYESIAN FORMULATIONS

Here in Chapter 8 we discuss Bayesian methods, both hierarchical and empirical, for defining the form and extent of shrinkage of sample estimates towards a subjective prior distribution. We find some striking parallels between Bayesian and classical shrinkage methodologies; formulas for point estimates of regression coefficients can frequently be made to agree exactly. But we also find profound differences; the Bayesian posterior variance of a point estimate is larger than the classical variance of that same estimate. We discuss not only why this type of difference exists but also point out some specific implications for statistical inference.

8.1 Bayesian Conjugate-Normal Linear-Model Formulations

Lindley and Smith(1972) describe a Bayesian formalism for hierarchical (multi-stage) analyses of linear models using conjugate multivariate-normal prior distributions. This formalism expresses unknown parameters at each stage of an analysis in terms of a linear model at the previous, lower stage. But, although dispersion matrices at each stage can be arbitrary, they must be known. And, at the final stage, both the mean vector and the dispersion matrix must be known. Here, we will be more interested in very simple, 2-stage analyses than in 3-or-more-stage (priors-on-priors) models. Thus our discussion will only rarely dwell deeper into Bayes' theory/practice than what is provided by "classic" reference works such as those of Raiffa and Schlaifer(1961) and Box and Tiao(1973).

The notation $y \sim N(\mu , D)$ will mean here that the column vector y has a multivariate normal distribution with mean given by the column vector μ and variance-covariance matrix given by the positive semi-definite matrix D. Similarly, $y \mid \theta \sim$ will mean that the conditional distribution of y given θ is being defined. In this notation, the fundamental lemma of Lindley and Smith(1972), pages 4-5, states that:

> LEMMA: If the sampling distribution of the response, y, is $y \mid \theta_1 \sim N(A_1 \theta_1 , D_1)$, where θ_1 is a $p_1 \times 1$ parameter vector, and the prior distribution is $\theta_1 \mid \theta_2 \sim N(A_2 \theta_2 , D_2)$ where θ_2 is a $p_2 \times 1$ parameter vector, then the marginal (unconditional) distribution of y is
>
> $$y \sim N(A_1 A_2 \theta_2 , D_1 + A_1 D_2 A_1^T) \qquad \{ 8.1 \}$$
>
> and the posterior (conditional) distribution of θ_1 given y is
>
> $$\theta_1 \mid y \sim N(B b , B) \qquad \{ 8.2 \}$$

where $B^{-1} = A_1^T D_1^{-1} A_1 + D_2^{-1}$ and $b = A_1^T D_1^{-1} y + D_2^{-1} A_2 \theta_2$.

To apply this lemma and demonstrate that a simple 2-stage Bayesian formalism produces generalized shrinkage estimators, we first make the identifications $A_1 \theta_1 = X \beta$ and $D_1 = \sigma^2 \cdot I$. Thus we are using a "point prior" on the error variance (i.e. proceeding as if σ^2 where known) and $B^{-1} = \sigma^{-2} \cdot X^T X + D_2^{-1}$. Next, we set the prior mean value for β to ZERO by taking $\theta_2 = 0$ and assure that D_2^{-1} (and D_2) will be simultaneously diagonalizable with $X^T X$ by restricting attention to prior variance-covariance matrices of the general form $D_2 = \sigma^2 \cdot G K^{-1} G^T$, where K is a diagonal R × R matrix and G is the P × R semi-orthogonal matrix of direction cosines for the principal axes of X , as in equation { 2.8 }. Now the Bayes point estimate is the mean, B b , of the posterior distribution of β given y (as well as given X) and this mean vector is of the general form:

$$E (\beta | y , X) = G (\Lambda+K)^{-1} \Lambda G^T y = G \Delta c = b^{\bigstar} \qquad \{ 8.3 \}$$

as in { 3.1 } , where $\Delta = \Lambda (\Lambda + K)^{-1} = K^{-1} (\Lambda^{-1} + K^{-1})^{-1}$ is the diagonal matrix of generalized shrinkage factors and c is the vector of uncorrelated components of the least-squares estimator. In other words, we have now demonstrated that all generalized shrinkage estimators are 2-stage Bayes estimates. This includes, of course, the special case of shrinkage estimates in the 2-parameter shrinkage family of { 3.9 } where $K = k \cdot \Lambda^Q$, the scalar Q determines the shape of the shrinkage path, and the scalar k (or, equivalently, $MCAL = R - \delta_1 - \ ... - \delta_R$) determines the extent of shrinkage. [Bayes estimates of more general form than { 8.3 } can, of course, result from choices of $\theta_2 \neq 0$ and/or $D_2 \neq \sigma^2 \cdot G K^{-1} G^T$.]

The above observation goes a long way, perhaps, towards explaining why many people apparently think that shrinkage estimation methods are Bayesian. But, wait a second! The Bayesian variance-covariance matrix, B , of the posterior distribution of the regression coefficient vector, β , given the observed vector of responses, y , (as well as given X) is of the general form:

$$V (\beta | y , X) = \sigma^2 \cdot (X^T X + G K G^T)^{-1} = \sigma^2 \cdot G \Delta \Lambda^{-1} G^T . \qquad \{ 8.4 \}$$

Note that the implied Bayesian dispersion (variance, co-variance) matrix for $G^T \beta = \gamma$ is the diagonal matrix $\sigma^2 \Delta \Lambda^{-1}$ and that this matrix has the same functional form as the classical, fixed-effect minimum risk, { 4.7 }, in $G^T b^{\bigstar} = \Delta \gamma$, which is achieved only when $\Delta = \Delta^{MSE}$. Note, in particular, the Bayesian dispersion is usually larger than the classical dispersion of equation { 3.4 },

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

After all, $\Delta^2 \leq \Delta$ when shrinkage factors are restricted to their "usual" range of $0 \leq \delta_i < 1$ for $1 \leq i \leq R$. In fact, strict in-equality ($\Delta^2 < \Delta$) holds whenever none of the shrinkage factors

is an actual ZERO. As we stress below, this difference in dispersion matrices has profound, practical implications for statistical inference.

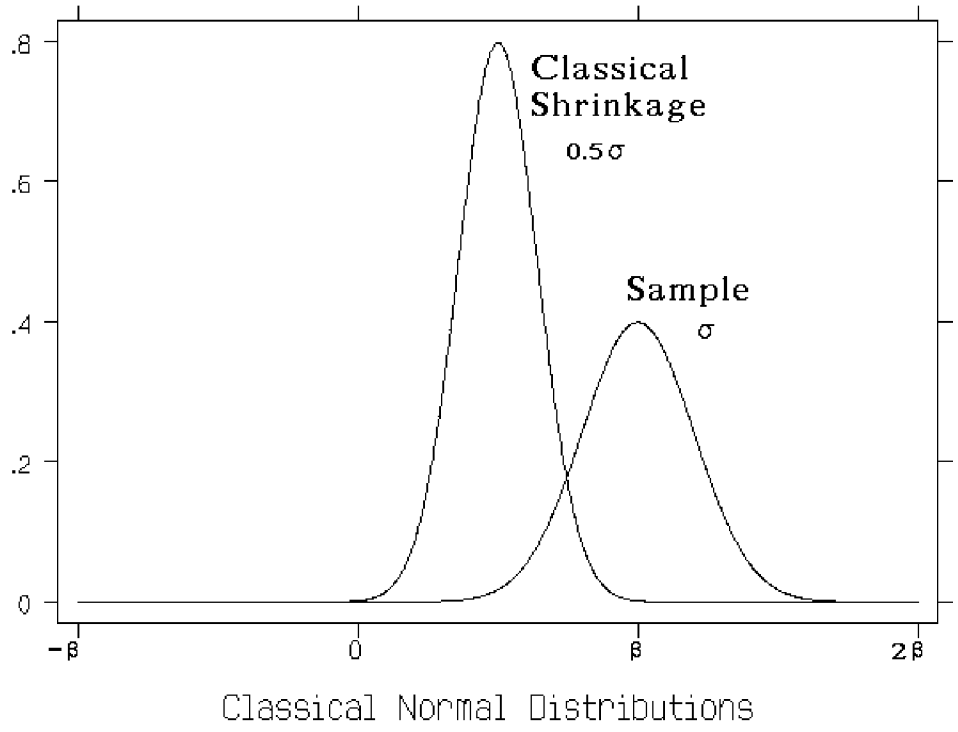
> Before starting our arguments on why Bayesian variances exceed their classical counterparts, we comment that substitution of estimates for unknown parameters into formulas { 8.3 } and { 8.4 } is commonly known as the "naive" empirical Bayes approach. This approach is apparently said to be naive because, relative to a full-blown hierarchical Bayes analysis, variances are thereby under estimated! For example, Ghosh(1992), pages 153-154, discusses an analysis that illustrates how naive substitutions ignore the "uncertainty involved in estimating the prior parameters when estimating the posterior variance." Then Ghosh(1992) argues [pages 168-173] that, although the empirical Bayes method of Morris(1983) is "an attempt to approximate a bonafide hierarchical Bayes procedure, and is clearly superior to a naive empirical Bayes procedure", variances are then over estimated by 11% in one example (and might be as much as 30% too large.) All that I really wish to stress here is that equation { 8.4 } apparently represents some sort of lower-bound for the variance of the shrinkage estimator { 8.3 } from Bayesian points-of-view. And yet this minimum Bayesian variance can still be considerably larger that the classical variance of that same estimator. Here's why...

Bayesian estimators incorporate added information from the prior distribution into the analysis. In fact, Bayes estimates are considered to be unbiased relative to combined sample and prior information about β. In other words, the variance-covariance matrix of a Bayes estimate is also its mean-squared-error matrix! In particular, the rank 1 squared-biases matrix of the classical formulation, the $(I - \Delta)\gamma\gamma^{T}(I - \Delta)$ term in { 4.2 }, is absent from the Bayesian formulation. And every choice for Δ yields Bayes risks for true components, γ, that behave like the minimum classical risks in Δ c achieved only at $\Delta = \Delta^{MSE}$.

From a Bayesian point-of-view, the more drastic is the shrinkage (the smaller is the Δ) imposed by a highly "informative" prior, the better-off one ends-up being! In other words, conflict between prior and sample information can be tolerated because "shrinkage" will effect a compromise. The more distinct/remote is the prior distribution from the sampling distribution, the more distinct/remote will be the posterior estimate from the sample estimate. In fact, one's prior distribution is more informative in these large-separation cases, and the Bayes posterior estimate ends up being correspondingly more precise.

Classical fixed-effect analyses of shrinkage estimators assume that bias is being introduced into the analysis. The multiplicative δ-factors in classical shrinkage formulas enter variance formulas as δ^2-factors. In other words, the standard deviations (square roots of variances) of classical shrinkage estimators are multiplied by the same δ-factors as are expected values; expected values and standard deviations thus change at exactly the same rate in classical shrinkage analyses. This point is illustrated in Figure 8.1 where a shrinkage factor of δ=0.5 changes the expected value of an estimate from β to $\beta/2$ and its standard deviation from σ to $\sigma/2$.

Figure 8.1 The Classical Shrinkage Formulation



Classical Normal Distributions

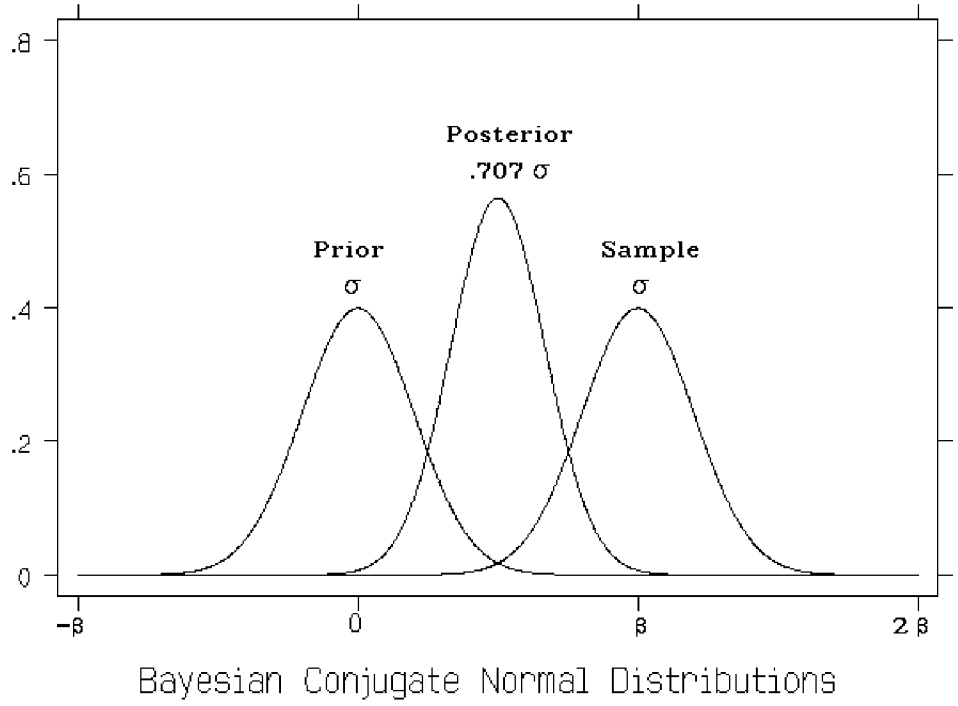
Shrinkage has no effect whatsoever on classical fixed-effect statistical inferences for regression coefficients that are based upon F-ratios and t-statistics. After all, a t-statistic is a ratio with an unbiased estimate of the numerical size of an effect in its numerator and a root-mean-square estimate of the corresponding standard deviation in its denominator; a F-ratio is the square of a t-statistic (or a sum-of-squares of several t-statistics with the same denominator). Anyway, there are convincing arguments [see Obenchain(1977) for details] that these ratios are actually invariant under classical shrinkage. In other words, classical fixed-effect shrinkage does not produce confidence intervals/regions for regression coefficients that are shifted in location and/or differnet in size from those derived by ordinary least-squares theory.

ASIDE: One might consider forming a confidence set for δ times β. Relative to a classical confidence set for β, the corresponding classical confidence set for $\delta \cdot \beta$ would be shifted in location (being centered at $\delta \cdot b^{o}$) and would be smaller in size (being based on dispersion $\delta \cdot s$) whenever $0 \leq \delta < 1$. On the other hand, confidence sets for $\delta \cdot \beta$ are of relatively little practical interest compared with the confidence set for the full β vector!

In summary, classical fixed-effect shrinkage methods are best applied on a contingency basis. The data at hand may provide convincing evidence of reduced dispersion that will more than offset the introduction of squared-bias, yielding an overall reduction in mean-squared-error. However, although one has the option of shrinking classical point estimates of regression coefficients, their corresponding fixed-effect set estimates remain unchanged. On the other hand, point and set estimates usually would change or shift, at least a little, if fixed-effects in a classical model were declared random. After all, BLUEs are then replaced with shrunken

BLUPs and confidence intervals/regions are then constructed using variance-component estimates!

Figure 8.2 A Bayesian Shrinkage Formulation



Because Bayesian posterior variances decrease in direct proportion to their δ shrinkage factors, Bayesian standard deviations decrease at a slower ($\delta^{1/2}$) rate than do their mean values. This point is illustrated in Figure 8.2, above, where a Bayes shrinkage factor of δ=0.5 results because the sample distribution (centered at β) and the prior distribution (centered at 0) are of exactly equal precision, σ. This Bayesian shrinkage produces a posterior distribution with expected value $\beta/2$, but the posterior standard deviation is $\sigma/\sqrt{2} = 0.707\ \sigma$ rather than $\sigma/2$.

Bayesian formulas for posterior F-ratios and t-statistics that measure differences between a posterior estimate and its prior mean tend to be “shrunken" in the sense that their numerators (effect sizes) have decreased more than their denominators (uncertainty measures.) This, of course, weakens any evidence that the posterior estimate might be discordant with the prior mean. In fact, Bayesian highest-posterior-density intervals/regions resulting from an informative prior for regression coefficients definitely are shifted in location (towards the prior) and are smaller in size than are the corresponding classical (frequentist) intervals/regions. By incorporating added information from the prior into the analysis, Bayes procedures end up “shrinking" highest-posterior-density set estimates as well as point estimates of regression coefficients.

8.2 Bayesian Diagnostic Checking

Informal methods for determining the effects of changes in one's Bayesian prior distribution upon the implied posterior distribution are commonly called "sensitivity analyses," Winkler(1972). Modern hardware/software reduces the implied computational burden to the point where at least some sort of diagnostic checking would seem to be a mandatory component of even in the most routine of Bayesian analyses. Box and Tiao(1973) and Berger(1980a, 1980b, 1983) suggest some more formal methods under the general title of Bayesian "robustness." And uncertainty about one's prior apparently motivates the 3rd-and-higher stages in the hierarchical approach of Lindley and Smith(1972).

In his "model adequacy" approach to "assessing the prior" that yields shrinkage regression estimates [Box(1980), Section §3], Box considers "predictive checks" derived using the marginal distribution, { 8.1 }. [This marginal distribution can be called "predictive" in the sense that it describes the expected behavior of sample data for the current analysis, but this marginal distribution is definitely distinct from the "predictive" distribution of a future sample, Zellner and Chetty(1965) or Aitchison and Dunsmore(1975), that results from integrating the sampling distribution over the posterior distribution.] In any case, we note that the mean and variance of the marginal distribution are:

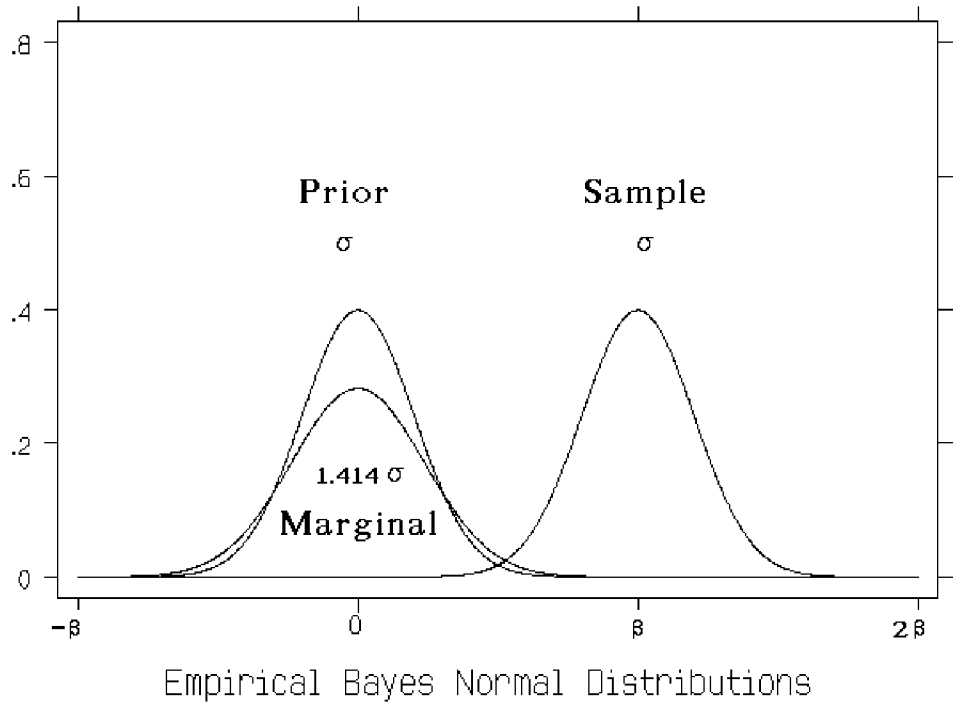
$$E(\beta \mid X) = 0 \quad \text{(the prior mean)} \qquad \{ 8.5 \}$$

and

$$V(\beta \mid X) = \sigma^2 \cdot G(I - \Delta)^{-1} \Lambda^{-1} G^T . \qquad \{ 8.6 \}$$

Thus, relative to the distribution of sample estimates, the marginal distribution is not only shifted in location (to the point that it totally ignores sample information!) but also has increased dispersion. These points are illustrated in Figure 8.3 where a Bayes shrinkage factor of δ=0.5 again results because the sample distribution (centered at β) and the prior distribution (centered at 0) are of exactly equal precision, σ. Note that the marginal distribution also has an increased standard deviation of $\sqrt{2}\,\sigma = 1.414\,\sigma$.

Figure 8.3 A Bayesian Marginal Distribution



Now Box(1980), equations (3.1) to (3.10), points out that his “predictive check" agrees with the Theil(1963) measure of the “compatibility of prior and sample information" and is defined as follows: For the extent of shrinkage implied by a given set of factors, Δ , calculate the F-ratio that measures the squared-distance between the least-squares estimates vector and the marginal mean in the metric of the marginal dispersion, namely

$$\text{Bayes predictive F-ratio} \;=\; c^T (I - \Delta) \Lambda c / (s^2 R), \qquad \{ 8.7 \}$$

where s^2 is the sample residual-mean-square. Then calculate the observed significance level of this F-ratio, which is the probability that a random variate with a central F-distribution (with R numerator degrees-of-freedom and N – R – 1 denominator degrees-of-freedom) would exceed that observed F value. This observed significance level allows any choice for the extent of shrinkage, Δ , to be “criticized."

It seems to me, at least, that the corresponding classical statistic would measure the squared-distance between the least-squares estimates and the shrunken estimates in the metric of the sampling dispersion, namely

$$\text{Classical F-ratio} \;=\; c^T (I - \Delta)^2 \Lambda c / (s^2 R). \qquad \{ 8.8 \}$$

Note that, just as Δ versus Δ^2 provide the distinction between the Bayes and classical dispersion matrices of { 8.4 } and { 3.4 }, $(I - \Delta)$ versus $(I - \Delta)^2$ provide the distinction between the Bayes and classical statistics of equations { 8.7 } and { 8.8 }. By the way, Obenchain(1977) called the observed significance level of { 8.8 } the associated probability of classical ridge shrinkage, while McCabe(1978) termed this same quantity the α-acceptability for that extent of shrinkage. Also, note that all significance levels (classical and Bayesian) are being computed relative to the same central F-distribution (with R numerator degrees-of-freedom and $N - R - 1$ denominator degrees-of-freedom.)

Next, note that there is a consistent difference between the Bayesian and classical observed significance levels associated with a given extent of shrinkage. Because the classical F-ratio of { 8.8 } is almost always smaller, numerically, than is the Bayes' predictive F-ratio of { 8.7 }, its significance level is almost always larger (less significant) than that given by the Bayesian evaluation of the same shrinkage. In other words, once a Bayesian becomes "introspective" about his/her specific choice of location and/or spread for a prior distribution, he/she is almost always more critical of his/her own choice than a classicist would be when evaluating the exact same form and extent of shrinkage.

8.3 More Bayes' Measures of the Extent of Shrinkage

There are at least two ways to quantify the extent of shrinkage employed in a given set of Bayes estimates for linear model coefficients.

Theil(1963) describes a variety of ideas, many of which were expanded on by later authors. For example, Theil(1963) proposes the "f-class" of mixed (empirical Bayes) estimators for regression, which provide a form of generalized shrinkage towards an origin space, and suggests (page 404) that these estimates be plotted as in a ridge TRACE type of display. And, as observed earlier, Theil(1963), equation (3.3), describes a special case of equation { 8.7 }. However, in my opinion, the primary contribution of Theil(1963) is his demonstration of uniqueness properties of a certain Bayesian measure of extent of shrinkage originally introduced by Schlaifer.

Theil started by asking a question like... "What proportions of posterior relative precision in a Bayes estimate are due, respectively, to sample information and to prior information." This is like asking... "What are the relative contributions of matrices A and B to the matrix $(A + B)^{-1}$?" Specifically, suppose that a function $g(A , B)$ is to be our measure the contribution of A to $(A + B)^{-1}$. Theil(1963) argued that the following four requirements on $g(A , B)$ seem reasonable.

(i) Adding-Up Criterion: $g(A , B) + g(B , A) \equiv 1$.

(ii) Zero Unit Criterion: $g(0 , B) = 0$ when $B \neq 0$

and

$g(A , 0) = 1$ when $A \neq 0$.

(iii) Invariance Under Nonsingular Linear Transformations, K , of Predictors:

$$g(K^T A K , K^T B K) \equiv g(A , B) .$$

(iv) Linearity Criterion: If A_1 , B_1 , A_2 and B_2 are such that $A_1 + B_1 = A_2 + B_2$ and p and q are two non-negative scalars that sum to one, then

$$g(p \cdot A_1 + q \cdot A_2 , p \cdot B_1 + q \cdot B_2) \equiv p \cdot g(A_1 , B_1) + q \cdot g(A_2 , B_2) .$$

Then Theil(1963) demonstrated that the unique measure satisfying all four of the above criteria is

$$g(A , B) = \text{trace}[A (A + B)^{-1}] / R , \qquad \{ 8.9 \}$$

when A and B are R × R matrices. Applying this result to the Bayes posterior variance, { 8.4 }, where the sampling precision is $A = \sigma^{-2} \cdot G \Lambda G^T$ and the prior precision is $B = \sigma^{-2} \cdot G K G^T$, we find that the proportion of posterior precision due to sample information is

$$g(\Lambda , K) = \text{trace}[\Lambda (\Lambda + K)^{-1}] / R = \sum \delta_i / R , \qquad \{ 8.10 \}$$

which is (R − MCAL) / R. In other words, when the multicollinearity allowance is MCAL = 0 [so that $\Delta = I$, and no shrinkage gets applied], then all posterior precision derives from sample information. But, at the other extreme of MCAL = R [where $\Delta = 0$, and total shrinkage to the prior mean is enforced], then none of posterior precision is derived from sample information.

Lindley(1980) writes that the "only satisfactory inference definition of information is surely Shannon's" and gives a formula that can be derived as follows: The Bayes estimator (posterior mean) of equation { 8.3 } can be written as a linear transformation, $b^{\bigstar} = (I + Z)^{-1} b^o$, of the sample (least-squares) estimator, b^o, where $Z = G \Lambda^{-1} K G^T$. Now Shannon's measure of information gain, posterior minus prior, is given by the corresponding difference in the expected values of the log likelihoods. For an R-dimensional multivariate normal distribution with dispersion matrix Σ , the expected log likelihood is $E(\ln L) = -\frac{1}{2} \cdot [R \cdot \ln(2\pi) + R + \ln |\Sigma|]$. As a result, Shannon's measure of information gain can be written as

$$\Im = \tfrac{1}{2} \cdot \ln[| I + Z | / | Z |] = -\tfrac{1}{2} \cdot \sum \ln(1 - \delta_i) . \qquad \{ 8.11 \}$$

Thus Shannon's measure of information gain is $\Im = +\infty$ when $\Delta = I$ [because the prior suggests no shrinkage whatsoever in this extreme case] and $\Im = 0$ when $\Delta = 0$ [because the posterior and prior coincide in this extreme case.]

8.4 Nonconjugate Bayes Formulations

I would like to make a couple of observations about nonconjugate Bayes analyses of linear models even though I am not sufficiently familiar with this literature to critique it here in any real detail. The nonconjugate analyses proposed by Draper and Van Nostrand(1977b) and Berger(1980b,1983) yield regression coefficient estimates of specifically nonlinear form. Their resulting risk (mean-squared-error) matrices strike me as being potentially more realistic than { 8.4 }. Unfortunately, the equations that define these sorts of estimates are sufficiently complicated, mathematically, that nothing short of detailed computational experience in applying these techniques to a wide variety of numerical examples would be adequate to appreciate how well they might perform in actual practice.

8.5 An Empirical Bayes Likelihood Approach

In his rejoinder to the discussion of his paper, Morris(1983) observes

> "Several discussants have gathered that `empirical Bayes' means plugging non-Bayesian estimates of the prior distribution into Bayes rules. I believe nothing in the empirical Bayes paradigm, or in frequency theory for that matter, forbids use of Bayes rules."

I agree and would add the thoughts: Bayes theorem is, after all, a theorem in classical statistics. Why should anybody feel hesitant to apply these tools in ways that they feel are appropriate and reasonable?

The empirical Bayes minus-two-log-likelihood factor of Efron and Morris(1977) for evaluating the extent of shrinkage is also based upon the marginal (predictive) distribution of equations { 8.1 }, { 8.5 }, and { 8.6 }. The minus-twice-log-likelihood resulting from treating the vector of least squares estimates for regression coefficients as if it were an observation from this marginal distribution [using the residual-mean-square s^2 as one's estimate of σ^2] is

$$-2 \cdot \ln(\mathrm{ML}) = R \cdot \ln(2 \cdot \pi \cdot s^2) + \sum_{i=1}^{R} \{ F_i \cdot (1 - \delta_i) - \ln[\lambda_i \cdot (1 - \delta_i)] \}, \qquad \{ 8.12 \}$$

where $F_i = c_i^2 \cdot \lambda_i / s^2$ is again the F-ratio of equation { 2.22 } for testing the statistical significance of the i-th uncorrelated component of the least-squares vector. When actually applying this criterion, Efron and Morris(1977) suggest simply ignoring all of the terms in { 8.12 } that do not change as shrinkage occurs. The factor they suggest computing is thus

$$\mathrm{EBAY} = \sum_{i=1}^{R} F_i \cdot (1 - \delta_i) + 2 \cdot \Im, \qquad \{ 8.13 \}$$

where $\Im = -\frac{1}{2} \cdot \sum \ln(1 - \delta_i)$ is Shannon's measure of information gain from equation { 8.11 }.

References for Chapter Eight

Aitchison, J. and Dunsmore, I. R. (1975). **Statistical Prediction Analysis**. Cambridge University Press.

Berger, J. O. (1980a). **Statistical Decision Theory: Foundations, Concepts, and Methods**. New York: Springer-Verlag.

Berger, J. O. (1980b). "A robust generalized Bayes estimator and confidence region for a multivariate normal mean." **Annals of Statistics** 8, 716-761.

Berger, J. O. (1983). "The robust bayesian viewpoint." **Robustness in Bayesian Statistics**. J. Kadane, ed. Amsterdam: North Holland.

Box, G. E. P. (1980). "Sampling and Bayes' inference in scientific modeling and robustness." **Journal of the Royal Statistical Society** A143, 383-404. (with discussion 404-430.)

Box, G. E. P. and Tiao, G. C. (1973). **Bayesian Inference in Statistical Analysis.** Reading, Massachusetts: Addison-Wesley.

Draper, N. R. and Van Nostrand, R. C. (1977a). "Ridge regression and James-Stein estimation: review and comments." **Technometrics** 21, 451-466.

Draper, N. R. and Van Nostrand, R. C. (1977b). "Ridge regression: is it worthwhile?" Technical Report No. 501, Department of Statistics, University of Wisconsin.

Efron B. and Morris, C. N. (1973). "Stein's estimation rule and its competitors." **Journal of the American Statistical Association** 68, 117-130.

Efron B. and Morris, C. N. (1975). "Data analysis using Stein's estimator and its generalizations." **Journal of the American Statistical Association** 70, 311-319.

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

Ghosh, M. (1992). "Hierarchical and empirical Bayes multivariate estimation." **Current Issues in Statistical Inference: Essays in Honor of D. Basu**. Institute of Mathematical Statistics, LECTURE NOTES-MONOGRAPH SERIES #17, 151-177.

Lindley, D. V. and Smith, A. F. M. (1972). "Bayes estimates for the linear model." **Journal of the Royal Statistical Society** B34, 1-72.

Lindley, D. V. (1980). "Comment" [on "A critique of some ridge methods" by Smith and Campbell.] **Journal of the American Statistical Association** 75, 94-95.

McCabe, G. P. (1978). "Evaluation of regression coefficients using α-acceptability." **Technometrics** 20, 131-139.

Morris, C. N. (1983). "Parametric empirical Bayes inference: theory and applications" **Journal of the American Statistical Association** 78, 47-55. (with discussion, 55-65.)

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

Obenchain, R. L. (1981). "Maximum likelihood ridge regression and the shrinkage pattern hypotheses." Abstract 81t-23. **Institute of Mathematical Statistics Bulletin** 10, 37.

Raiffa, H. and Schlaifer, R. (1961). **Applied Statistical Decision Theory**. Harvard University Press.

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

Rolph, J. E. (1976). "Choosing shrinkage estimators for regression problems." **Communications in Statistics** A5, 789-802.

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

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

Winkler, R. L. (1972). **An Introduction to Bayesian Inference and Decision**. New York: Holt, Rinehart and Winston.

Zellner, A. and Chetty, V. K. (1965). "Prediction and decision problems in regression models from the Bayesian point of view." **Journal of the American Statistical Association** 60, 608-616.

Further Reading for Chapter Eight

Anderson, R. L. and Battiste, E. L. (1975). "The use of prior information in linear regression analysis." **Communications in Statistics** 4, 497-517.

Bacon, R. W. and Hausman, J. A. (1974). "The relationship between ridge regression and the minimum mean square error estimator of Chipman." **Oxford Bulletin of Economics and Statistics**, 36, 115-124.

Banerjee, K. S. and Carr, R. N. (1971). "A comment on ridge regression, biased estimation for nonorthogonal problems." **Technometrics** 13, 895-898.

Brown, P. and Payne C. (1975). "Election night forecasting." **Journal Royal Statistical Society** A138, 463-498.

Guilkey, D. K. and Murphy, J. L. (1975). "Directed ridge regression techniques in cases of multicollinearity." **Journal American Statistical Association** 70, 769-775.

Hsiang, T. C. (1976). "A Bayesian view on ridge regression." **The Statistician** 24, 267-268.

Goldstein, M. and Smith, A. F. M. (1974). "Ridge-type estimators for regression analysis." **Journal Royal Statistical Society** B36, 284-291.

Goldstein, M. (1976). "Bayesian analysis of regression problems." **Biometrika** 63, 51-58.

Good, I. J. (1965). **The Estimation of Probabilities, An Essay on Modern Bayesian Methods.** Cambridge, Massachusetts: M.I.T. Press.

Leamer, E. E. (1977). "Valley regression: biased estimation for orthogonal problems." Department of Economics, University of California, Los Angeles.

Leamer, E. E. (1978). "Regression selection strategies and revealed priors." **Journal of the American Statistical Association** 73, 580-587.

Obenchain, R. and Vinod, H. (1974). "Estimates of partial derivatives from ridge regression on ill-conditioned data." **NBER-NSF Seminar on Bayesian Inference in Econometrics**, Ann-Arbor, Michigan.

Smith, A. F. M. and Goldstein, M. (1975). "Ridge regression: some comments on a paper of Conniffe and Stone." **The Statistician**, 24, 61-66.

Stone, J. and Conniffe, D. (1973). "A critical view of ridge regression." **The Statistician**, 22, 181-187.

Swamy, P. A. V. B., Rappoport Paul N. (1975). “Relative efficiencies of some simple Bayes estimators of coefficients in dynamic models - I." **Journal of Econometrics** 3, 273-296. 1976.