CHAPTER 17

Ridge Regression

17.1. INTRODUCTION

The *ridge trace* procedure, first suggested by A. E. Hoerl in 1962, is discussed at length in two seminal papers by Hoerl and Kennard (1970a,b). The procedure is intended to overcome "ill-conditioned" situations where near dependencies between various columns in **X** cause the **X'X** matrix to be close to singular, giving rise to unstable parameter estimates, typically with large standard errors. Because ill-conditioning is a function of model and data, it would usually be wise to find out how the ill-conditioning occurs, as discussed in Chapter 16. Ridge regression is a way of proceeding that adds specific additional information to the problem to remove the ill-conditioning. Not all users realize this and ridge regression is sometimes used when it is entirely inappropriate. This is not the fault of the originators, who applied the technique to industrial problems in sensible ways.

17.2. BASIC FORM OF RIDGE REGRESSION

The ridge regression procedure, in its simplest form, is as follows. Let **F** represent the appropriate centered and scaled "**X** matrix" when the regression problem under study is in "correlation form" (as discussed in Chapter 16). Thus, if the original model is

$$Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \cdots + \beta_r Z_r + \epsilon, \tag{17.2.1}$$

the new centered and scaled predictor variables are

$$f_{ju} = \frac{Z_{ju} - \overline{Z}_j}{S_{jj}^{1/2}}, \tag{17.2.2}$$

where $\overline{Z}_j$ is the average of the Z_{ju}, $u = 1, 2, \ldots, n$, and $S_{jj} = \Sigma_u (Z_{ju} - \overline{Z}_j)^2$. Thus

$$\mathbf{F} = \begin{bmatrix} f_{11} & f_{21} & \cdots & f_{r1} \\ \vdots & \vdots & & \vdots \\ f_{1u} & f_{2u} & \cdots & f_{ru} \\ \vdots & \vdots & & \vdots \\ f_{1n} & f_{2n} & \cdots & f_{rn} \end{bmatrix} \tag{17.2.3}$$

and **F'F** is the correlation matrix of the Z's. The size of **F'F** is r by r; the original

model had $(r + 1)$ parameters in it, including β_0. Note that $\mathbf{F'1} = \mathbf{0}$, because of the centering.

In Chapter 16 we also put $\mathbf{Y}$ in centered and scaled form. Usually $\mathbf{Y}$ is only centered in ridge regression, but it could be scaled, too, if desired. If $\mathbf{Y}$ is not scaled, the least squares model (17.2.1) can be fitted as

$$Y_u - \overline{Y} = \beta_{1F} f_{1u} + \beta_{2F} f_{2u} + \cdots + \beta_{rF} f_{ru} + \epsilon_u \tag{17.2.4}$$

with a vector of least squares estimates we shall refer to as

$$\mathbf{b}_F(0) = \{b_{1F}(0), b_{2F}(0), \ldots, b_{rF}(0)\}', \tag{17.2.5}$$

say. We can think of $\overline{Y}$ as $b_{0F}(0)$. The reason for the zeros in parentheses will become apparent in a moment. The ridge regression estimates of the r elements of $\boldsymbol{\beta}_F = (\beta_{1F}, \beta_{2F}, \ldots, \beta_{rF})'$ are the r elements of $\mathbf{b}_F(\theta) = \{b_{1F}(\theta), b_{2F}(\theta), \ldots, b_{rF}(\theta)\}'$ given by the equation

$$\mathbf{b}_F(\theta) = (\mathbf{F'F} + \theta \mathbf{I}_r)^{-1} \mathbf{F'Y}, \tag{17.2.6}$$

where θ is a positive number. [In applications, the interesting values of θ usually lie in the range (0, 1). An exception is described by Brown and Payne (1975).] Note that, when $\theta = 0$, we obtain the least squares estimates. The reader may wonder why $\mathbf{Y}$, rather than $\mathbf{Y} - \overline{Y}\mathbf{1}$ (the centered vector), appears on the right of (17.2.6). The reason is that $\overline{Y}\mathbf{1}$ makes no difference to the calculations at all; the centering of the predictor variables causes $\mathbf{F'1} = 0$, so that $\mathbf{F'}(\mathbf{Y} - \overline{Y}\mathbf{1}) = \mathbf{F'Y}$. To recover estimates of the original parameters in (17.2.1) we refer to (16.3.6) but remember that, because $\mathbf{Y}$ is centered but not scaled, the factor $S_{YY}^{1/2}$ is not needed. (It will be needed only if the Y's are centered *and* scaled.) The conversions are

$$b_j(\theta) = b_{jF}(\theta)/S_{jj}^{1/2}, \qquad j = 1, 2, \ldots, r, \tag{17.2.7}$$

and

$$b_0(\theta) = \overline{Y} - \sum_{j=1}^{r} b_j(\theta) \overline{Z}_j \tag{17.2.8}$$

to provide the $(r + 1) \times 1$ vector $\mathbf{b}(\theta) = \{b_0(\theta), b_1(\theta), \ldots, b_r(\theta)\}'$. When $\theta = 0$, the $b_j(0)$, $j = 0, 1, 2, \ldots, r$, are the usual least squares estimates for the parameters in (17.2.1), as is clear from setting $\theta = 0$ in Eqs. (17.2.6) to (17.2.8).

We can now plot the $b_{jF}(\theta)$ or the $b_j(\theta)$ against θ for $j = 1, 2, \ldots, r$ and examine the resulting figure; the intercept is usually not plotted. Such a plot is called a *ridge trace*. Typically this plot is drawn in "correlation" units [i.e., the $b_{jF}(\theta)$], to enable a direct comparison to be made between the relative effects of the various coefficients, and to remove the effects of the various Z scalings that affect the sizes of the original estimated coefficients. [In our example below, however, we have converted the plot to "original" units–that is, the $b_j(\theta)$—so that, when $\theta = 0$, we obtain the least squares estimates that would be obtained if the original Z's were *not* scaled.] As θ is increased to values beyond 1, the estimates typically become smaller in absolute value, and they tend to zero as θ tends to infinity. A value of θ, say, θ^*, is then selected. Hoerl and Kennard (1970a, p. 65) say:

> These kinds of things can be used to guide one to a choice:
>
> 1. At a certain value of θ the system will stabilize and have the general characteristics of an orthogonal system.

2. Coefficients will not have unreasonable absolute values with respect to the factors for which they represent rates of change.

3. Coefficients with apparently incorrect signs at $\theta = 0$ will have changed to have the proper signs.

4. The residual sum of squares will not have been inflated to an unreasonable value. It will not be large relative to the minimum residual sum of squares or large relative to what would be a reasonable variance for the process generating the data.

The first point must be interpreted in the context of small θ, $0 \leq \theta \leq 1$, of course, because as $\theta \to \infty$, all the $b_{jF} \to 0$. Points 2 and 3 require some prior knowledge for proper application. Point 4 raises the question of what is an unreasonable value. Fortunately, as we shall see in our example, an automatic (but sensible) way of choosing θ is available for those whose prior knowledge is insufficient.

Once θ has been selected (equal to θ^*) the values $b_j(\theta^*)$ are used in the prediction equation. The resulting equation is made up of estimates that are not least squares and are biased but that are more stable in the sense described and that (it is hoped; see Appendix 17B) provide a smaller overall mean square error, since the reduction they achieve in variance error will more than compensate for the bias introduced.

(Note that the biased estimates, which can be selected if desired through the Mallows C_p procedure, are biased by coefficients that have not been taken into account in the model fitted. The biased estimates in the Hoerl and Kennard procedure are biased by the presence of θ in the estimation equation and this bias will occur even if the postulated equation contains all the "right" predictor variables. In other words, the two sorts of biases are of different characters.)

17.3. RIDGE REGRESSION OF THE HALD DATA

We now apply this method to the Hald data, to illustrate the mechanics of the procedure. Because of the mixture relationship between the four predictor variables, these data could possibly give rise to unstable estimates, as we have already discussed.

Figure 17.1 shows a ridge trace for the Hald data for $0 \leq \theta \leq 1$ while Figure 17.2 shows an enlargement of the same figure for $0 \leq \theta \leq 0.03$. What value θ^* should be selected?

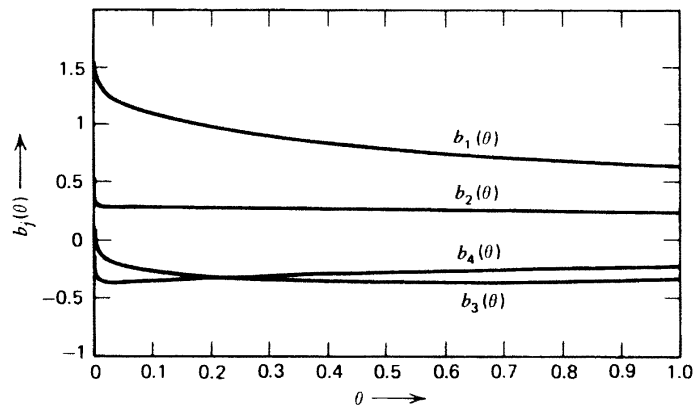

Figure 17.1. Ridge trace of Hald data, $0 \leq \theta \leq 1$. We are grateful to Dr. R. W. Kennard who kindly provided the results from which this figure was drawn.

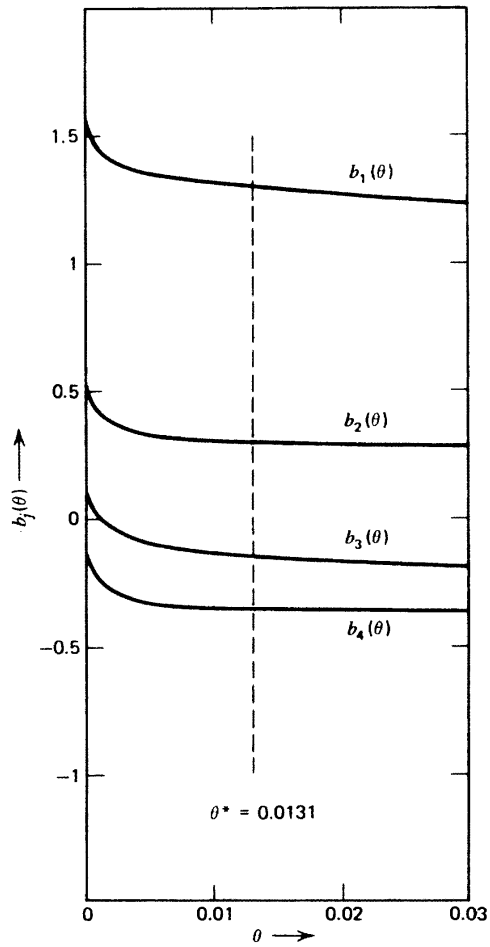

Figure 17.2. Ridge trace of Hald data, $0 \leq \theta \leq 0.03$. We are grateful to Dr. R. W. Kennard who kindly provided the results from which this figure was drawn.

Automatic Choice of θ^*

One suggested automatic way of choosing θ^* is given by Hoerl, Kennard, and Baldwin (1975). They argue that a reasonable choice might be

$$\theta^* = rs^2/\{\mathbf{b}_F(0)\}'\{\mathbf{b}_F(0)\}, \tag{17.3.1}$$

where

1. r is the number of parameters in the model *not counting* β_0.
2. s^2 is the residual mean square in the analysis of variance table obtained from the standard least squares fit of (17.2.1), or that of (17.2.4); the results are the same.
3. $$\begin{aligned}\{\mathbf{b}_F(0)\}' &= \{b_{1F}(0), b_{2F}(0), \ldots, b_{rF}(0)\} \\ &= \{\sqrt{S_{11}}b_1(0), \sqrt{S_{22}}b_2(0), \ldots, \sqrt{S_{rr}}b_r(0)\}.\end{aligned} \tag{17.3.2}$$
   (These two may differ slightly in practice due to rounding error.)

[Note that θ^* in Eq. (17.3.1) is s^2 (an estimate of σ^2) divided by the average value of the squares of the least squares estimates $b_{jF}(0)$. The latter can be regarded as an

estimate of σ_β^2, the variance of the true but unknown β_F values. It follows that, from a Bayesian viewpoint (see Section 17.4), the value of θ^* in (17.3.1) is a sensible choice.]

For these data, we have $r = 4$, $s^2 = 5.983$ (see Appendix 15A), and $\mathbf{b}_F(0) = (31.633, 27.516, 2.241, -8.338)'$, so that $\theta^* = 4(5.983)/1832.4 = 0.0131$. A vertical line is drawn at this θ-value in Figure 17.2, and the corresponding values of the coefficients can be read off the graph, or calculated more accurately. The following equation results, in fact:

$$\hat{Y} = 83.414 + 1.300X_1 + 0.300X_2 - 0.142X_3 - 0.349X_4. \tag{17.3.3}$$

This equation can be used in its present form.

Possible Use of Ridge Regression as a Selection Procedure; Other θ^*

Alternatively, consideration can be given to the possible deletion of one or more predictor variables. An obvious first choice is X_3. The value $b_3(0.0131) = -0.142$ is not only the smallest coefficient but, in addition, the values of X_3 in the data are such that the *maximum* effect on the response from X_3 is only $-0.142(23) = -3.266$. Probably the most sensible first steps to take would be to eliminate X_3 and then to refit and perform a new ridge trace for an equation containing X_1, X_2, and X_4. We do not pursue this matter here but refer the reader to the applications paper by Hoerl and Kennard (1970b) for comments on the considerations involved. It should be noted that, in applications, ridge regression has *not* usually been used as a selection-type procedure. We mention this use only as a possibility.

There are other ways of choosing θ^*. One, for example, is to use an iterative procedure. The basic idea is this: In the choice of θ^* above, a denominator of $\{\mathbf{b}_F(0)\}'\{\mathbf{b}_F(0)\}$ was used. For that reason let us call that value θ_0^*. Consider the iterative formula

$$\theta_{j+1}^* = rs^2[\mathbf{b}_F(\theta_j^*)\}'\{\mathbf{b}_F(\theta_j^*)\}]. \tag{17.3.4}$$

Use of θ_0^* on the right-hand side of this formula will enable us to obtain a revised value θ_1^*, which can then be substituted into the right-hand side of the formula to update to θ_2^*, and so on. Updating continues until

$$(\theta_{j+1}^* - \theta_j^*)/\theta_j^* \leq \delta, \tag{17.3.5}$$

where δ is appropriately chosen. For more on this procedure, see Hoerl and Kennard (1976). The authors note that $\theta_{j+1}^* \geq \theta_j^*$ always so that $\delta \geq 0$ and suggest

$$\delta = 20\{\text{trace}(\mathbf{F}'\mathbf{F})^{-1}/r\}^{-1.30} \tag{17.3.6}$$

as a suitable practical value, for reasons they describe on (their) pp. 79–80.

There is no mechanically best way of choosing θ^*. As we have already noted, Eq. (17.3.1) has some intuitive appeal because it can be regarded as a rough estimate of the ratio σ^2/σ_β^2 mentioned in the Bayesian formulation that follows. The iterative Eq. (17.3.4) can be regarded as giving another such estimate.

17.4. IN WHAT CIRCUMSTANCES IS RIDGE REGRESSION ABSOLUTELY THE CORRECT WAY TO PROCEED?

Opinions about the value of the ridge regression method vary over a large range. Before setting out our own opinions, we describe two (rather restrictive) situations

in which we can wholeheartedly recommend ridge regression as the best treatment of the problem described.

1. A Bayesian formulation of a regression problem with specific prior knowledge of a certain type on the parameters. This is discussed by Goldstein and Smith (1974, p. 291), and also mentioned in the original Hoerl and Kennard (1970a) paper. Essentially, ridge regression can be regarded as an estimate of $\boldsymbol{\beta}_F$ from the data, subject to the prior knowledge (or belief) that smaller values (in modulus, i.e., in size ignoring sign) of the β_{jF} are more likely than larger values, and that larger and larger values of the β_{jF} are more and more unlikely. More accurately, this prior knowledge would be expressed by imagining a multivariate normal distribution of prior belief on the values of the β_{jF}, a distribution whose spread would depend on a parameter σ_β^2. The parameter θ of the ridge regression would, in fact, be the ratio σ^2/σ_β^2 (where σ^2 is the usual variance of a single observation). Thus choice of a value of θ in ridge regression is equivalent to an expression of how big we believe the β_{jF} might be. Use of a very small θ (<0.01, say) would imply that we believed quite large values of the β_{jF} were not unreasonable. (The least squares estimates correspond to $\theta = 0$, or $\sigma_\beta^2 = \infty$, which means that, *a priori*, we do not feel we wish to restrict the β_{jF} at all.) Choice of a large θ (>1, say) would imply that we believed the β_{jF} were more likely to be quite small than otherwise. If we think of ridge regression in this way, then, we see that the "stabilization" of the ridge trace as θ increases is really *not* something implied by the data but is created by an *a priori* restriction of the possibilities for the parameter values.

2. A formulation of a regression problem as one of least squares subject to a specific type of restriction on the parameters. Suppose we perform least squares subject to the spherical restriction

$$\boldsymbol{\beta}_F'\boldsymbol{\beta}_F \leq c^2, \tag{17.4.1}$$

where c^2 is a specified value. If we use the method of Lagrange multipliers (see Appendix 9A) to do this, we can form the objective function

$$f \equiv (\mathbf{Y} - \bar{Y}\mathbf{1} - \mathbf{F}\boldsymbol{\beta}_F)'(\mathbf{Y} - \bar{Y}\mathbf{1} - \mathbf{F}\boldsymbol{\beta}_F) + \theta(\boldsymbol{\beta}_F'\boldsymbol{\beta}_F - c^2), \tag{17.4.2}$$

using the equality in the restriction, whereupon setting $\partial f/\partial \boldsymbol{\beta}_F = 0$ provides the equations

$$(\mathbf{F}'\mathbf{F} + \theta\mathbf{I})\boldsymbol{\beta}_F = \mathbf{F}'(\mathbf{Y} - \bar{Y}\mathbf{1}) = \mathbf{F}'\mathbf{Y}, \tag{17.4.3}$$

which leads to the ridge solution Eq. (17.2.6). This must be solved together with the assumed restriction $\boldsymbol{\beta}_F'\boldsymbol{\beta}_F = c^2$. If we solve for $\boldsymbol{\beta}_F$ in Eq. (17.4.3) and substitute into the restriction, we get

$$\mathbf{Y}'\mathbf{F}(\mathbf{F}'\mathbf{F} + \theta\mathbf{I}_r)^{-2}\mathbf{F}'\mathbf{Y} = c^2 \tag{17.4.4}$$

an equation that determines θ in terms of the given c value. Thus ridge regression can be viewed as least squares subjected to a spherical restriction on the parameters, the appropriate value of θ being determined by the radius c of the restriction. Figure 17.3 shows a geometric representation of this for the two-parameter (in addition to β_0) case. The usual (unrestricted) least squares solution occurs as the "bottom of the bowl" defined by the elliptical sum of squares contours. (An explanation of these is given in Chapter 24. We ask the reader who has not read those details to bear with us and gather a "general impression" of what is happening without worrying about the details here.) The spherical (here circular) restriction is shown and the restricted solution is on the innermost elliptical contour that just touches the spherical restriction. We obtain the whole sequence of ridge regression solutions as we "close in" the circle,

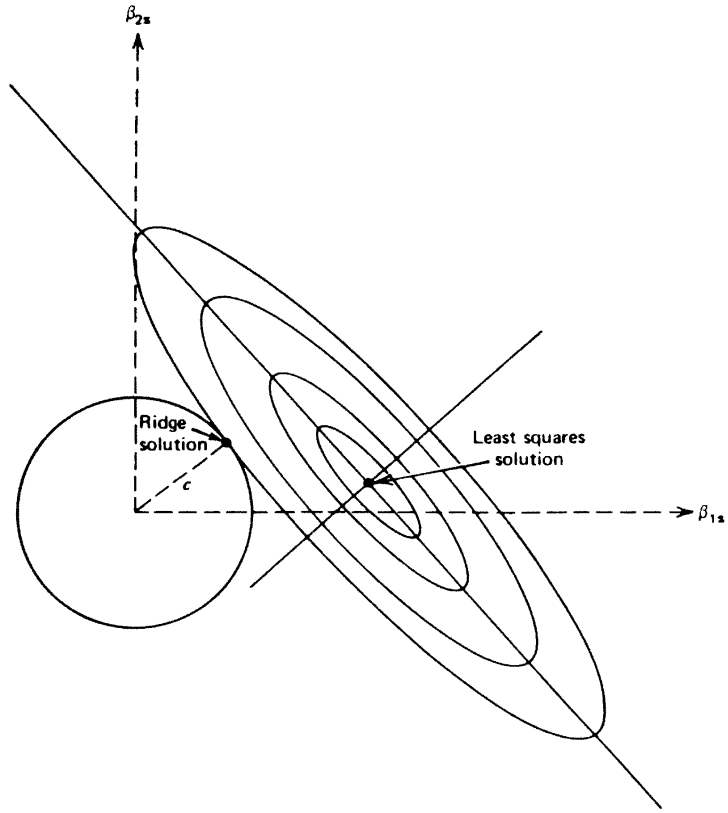

Figure 17.3. Ridge regression as the solution to a type of restricted least squares problem.

that is, reduce the radius from the length where the circle would just pass through the least squares solution (which would correspond to $\theta = 0$) to the "point circle" where $c = 0$ (which would correspond to $\theta = \infty$). Larger radius values would return us to the least squares solution, of course. The "ridge" of ridge regression is the path traced by the solution point as the circle radius is contracted or expanded from one extreme to the other of its effective range.

An alternative viewpoint is that ridge regression is the minimax solution to least squares when $\boldsymbol{\beta}'\boldsymbol{\beta} \leq \sigma^2/\theta$; see Bunke (1975).

Comments

We have said that situations (1) and (2) have the ridge regression formula as their solution. Provided we feel we can take one of these two approaches then, ridge regression is *exactly* the appropriate solution. There are two troubling aspects of this, however.

1. Can we really specify our prior knowledge in either of the ways given above? Are we *that* sure? For example, if we want to specify an ellipsoidal restriction

$$\boldsymbol{\beta}_F' \mathbf{U} \boldsymbol{\beta}_F \leq c^2 \tag{17.4.5}$$

instead of the spherical one, Eq. (17.4.1), we shall be led to a ridge-type solution of different form,

$$(\mathbf{F}'\mathbf{F} + \theta\mathbf{U})^{-1}\mathbf{F}'\mathbf{Y}, \tag{17.4.6}$$

instead. Thus the precise form of our restriction can be critical to the solution we obtain.

2. When we use the standard form [or any other form as in Eq. (17.4.6)] of ridge regression, we are—*whether we like it or not*—essentially making some sort of assumption of the type (1) or (2) given above, without really specifying it. The ridge regression solution is not a miraculous panacea. It is a least squares solution restricted by the addition of some external information about the parameters. The blind use of ridge regression without this realization can be dangerous and misleading. If the external information is sensible and is not contradicted by the data, then ridge regression is also sensible. A simple test of compatibility is to check whether the ridge estimates of the original parameters lie within, or on, the 95% (or other selected percentage) ellipsoidal confidence contour for those parameters. The contour is given by Eq. (5.3.7). When the appropriate ridge estimate vector is inserted in the positions occupied by $\boldsymbol{\beta}$, the left-hand side of (5.3.7) will be smaller than the right-hand side if the ridge estimate point falls inside the contour; equality means it falls on the contour, and left $>$ right in (5.3.7) means it falls outside the contour.

An additional worrying feature of ridge regression is the following. The characteristic effect of the ridge regression procedures is to change (from the least squares values) the nonsignificant estimated regression coefficients (whose values are statistically doubtful, anyway) to a far greater extent than the significant estimated coefficients. It is questionable that much real improvement in estimation can be achieved by such a procedure. (Most variable selection procedures, in the same context, would assign a value zero to a nonsignificant coefficient rather than adjusting it to a different but also nonsignificant value and such action would be no less reasonable and would have the virtue of concentrating attention on the apparently important predictors.)

Canonical Form of Ridge Regression

Much of the published technical work on ridge regression is done on the "canonical form," obtained by applying a transformation to convert the "$\mathbf{F}'\mathbf{F}$ matrix" into a diagonal matrix. This simplifies the algebra and provides a convenient way of making ridge trace calculations. The canonical form of ridge regression is discussed in Appendix 17C.

17.5. THE PHONEY DATA VIEWPOINT

A third interpretation of ridge regression is possible. We can introduce additional information into a regression problem by adding "prior data" of the form $\mathbf{F}_0$, $\mathbf{Y}_0$. If we give a weight of 1 to each real data point and a weight θ to each prior data point, the weighted least squares estimator of $\boldsymbol{\beta}_F$ is

$$\mathbf{b}_{FPD} = (\mathbf{F}'\mathbf{F} + \theta\mathbf{F}_0'\mathbf{F}_0)^{-1}(\mathbf{F}'\mathbf{Y} + \theta\mathbf{F}_0'\mathbf{Y}_0). \tag{17.5.1}$$

Suppose our prior data are "phoney" and are chosen so that $\mathbf{F}_0'\mathbf{F}_0 = \mathbf{I}$ and $\mathbf{Y}_0 = \mathbf{0}$; then we obtain the ridge estimator Eq. (17.2.6). Thus another viewpoint of ridge regression is as the tempering of the data by some zero observations on an orthogonal design of unit radius, appropriately weighted.

An alternative viewpoint is to choose $\mathbf{F}_0'\mathbf{F} = \theta\mathbf{I}$ for selected θ, $\mathbf{Y}_0 = \mathbf{0}$ and perform

an unweighted regression analysis. Thus readers who do not have a ridge regression package can obtain the same results by using a standard regression package and adding one dummy case for each of the "r" variables in the regression model with $\theta^{1/2}$ as the value for the corresponding variable and zeros for the rest of the variables in the model. In other words, $\mathbf{F}_0 = \theta^{1/2}\mathbf{I}_r$ and $\mathbf{Y}_0 = \mathbf{0}$.

In either event, we are adjoining some zero observations onto the data.

This method must be used with caution. The correct ridge regression coefficients are produced by it, but most of the other traditional computer output is no longer meaningful for the original data, providing a dangerous trap to the unwary. (On the other hand, it can be argued that this output correctly represents the effects of the assumption made.)

17.6. CONCLUDING REMARKS

Ridge Regression Simulations—A Caution

A number of papers containing simulations of regression situations claim to show that ridge regression estimates are better than least squares estimates when judged via mean square error. Such claims must be viewed with caution. Careful study typically reveals that the simulation has been done with effective restrictions on the true parameter values—precisely the situations where ridge regression is the appropriate technique theoretically. The extended inference that ridge regression is "always" better than least squares is, typically, completely unjustified.

Summary

From this discussion, we can see that use of ridge regression is perfectly sensible in circumstances in which it is believed that large β-values are unrealistic from a practical point of view. However, it must be realized that the choice of θ is essentially equivalent to an expression of how big one believes those β's to be. In circumstances where one cannot accept the idea of restrictions on the β's, ridge regression would be completely inappropriate.

Note that, in many sets of data, where the sizes of the least squares estimates are acceptable as they are, the ridge trace procedure would result in a choice of $\theta = 0$. A value of $\theta \neq 0$ would be used only when the least squares results were not regarded as satisfactory.

Opinion

Ridge regression is useful and completely appropriate in circumstances where it is believed that the values of the regression parameters are unlikely to be "large" (as interpreted through σ_β^2 or c^2 above). In viewing the ridge traces, a subjective judgment is needed that either (1) effectively specifies one's Bayesian prior beliefs as to the likely sizes of the parameters, or (2) effectively places a spherical restriction on the parameter space. The procedure is very easy to apply and a standard least squares regression program could easily be adapted by a skilled programmer. Overall, however, we would advise against the indiscriminate use of ridge regression unless its limitations are fully appreciated. (The reader should be aware that many writers disagree with our somewhat pessimistic assessment of ridge regression.)

17.7. REFERENCES

There is a great deal of published work on ridge regression and we give only some selected references. In the body of the text we have already mentioned Hoerl and Kennard (1970a,b, 1976); Hoerl, Kennard, and Baldwin (1975); Goldstein and Smith (1974); and Bunke (1975). For applications see Anderson and Scott (1974), Füle (1995), Marquardt (1970), and Marquardt and Snee (1975), which we especially recommend. For further work, see Coniffe and Stone (1973), Draper and Herzberg (1987), Draper and Van Nostrand (1979), Egerton and Laycock (1981), Gibbons (1981), Hoerl and Kennard (1975, 1981), Lawless and Wang (1976), McDonald (1980), Oman (1981), Park (1981), Smith (1980), and Swindel (1981). For ridge regression in other circumstances, see Chaturvedi (1993), Kozumi and Ohtani (1994), le Cessie and van Houwelingen (1992), Nyquist (1988), and Segerstedt (1992). For variations of ridge regression and links with other things, see Copas (1983); Crouse, Jin, and Hanumara (1995), Obenchain (1978); Panopoulos (1989); Saleh and Kibria (1993); Srivastava and Giles (1991); and Walker and Birch (1988).

APPENDIX 17A. RIDGE ESTIMATES IN TERMS OF LEAST SQUARES ESTIMATES

How do the ridge regression estimates relate to the least squares estimates? We recall from Eq. (17.2.6) that

$$\mathbf{b}_F(\theta) = (\mathbf{F}'\mathbf{F} + \theta\mathbf{I}_r)^{-1}\mathbf{F}'\mathbf{Y}. \tag{17A.1}$$

Using the fact that $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$, we can remove a factor $(\mathbf{F}'\mathbf{F})^{-1}$ from the right of the inverted matrix in (17A.1) to obtain

$$\begin{aligned} \mathbf{b}_F(\theta) &= \{\mathbf{I} + \theta(\mathbf{F}'\mathbf{F})^{-1}\}^{-1}\mathbf{b}_F(0) \\ &= \mathbf{T}\mathbf{b}_F(0) \\ &= \mathbf{T}\mathbf{b}_F, \end{aligned} \tag{17A.2}$$

say. Because

$$\mathbf{b}_F = \mathbf{b}_F(0) = \{b_{1F}(0), b_{2F}(0), \ldots, b_{rF}(0)\}' \tag{17A.3}$$

is the vector of least squares estimates after centering and scaling, except for the intercept term, we see that the ridge estimators are all linear combinations of the corresponding least squares estimators with coefficients determined by the matrix

$$\mathbf{T} = \{\mathbf{I} + \theta(\mathbf{F}'\mathbf{F})^{-1}\}^{-1}. \tag{17A.4}$$

APPENDIX 17B. MEAN SQUARE ERROR ARGUMENT

Ridge regression is sometimes "justified" as a practical technique by the claim that it produces lower mean square error. The basic result is as follows (e.g., Hoerl and Kennard, 1970a, p. 62). The mean square error of the ridge estimator can be written

$$\begin{aligned} \mathrm{MSE}(\theta) = E\{\mathbf{b}_F(\theta) - \boldsymbol{\beta}_F\}'\{\mathbf{b}_F(\theta) - \boldsymbol{\beta}_F\} &= E\{\mathbf{T}\mathbf{b}_F - \boldsymbol{\beta}_F\}'\{\mathbf{T}\mathbf{b}_F - \boldsymbol{\beta}_F\} \\ &= E\{(\mathbf{b}_F - \boldsymbol{\beta}_F)'\mathbf{T}'\mathbf{T}(\mathbf{b}_F - \boldsymbol{\beta}_F) \\ &\quad + \boldsymbol{\beta}_F'(\mathbf{T} - \mathbf{I})'(\mathbf{T} - \mathbf{I})\boldsymbol{\beta}_F. \end{aligned} \tag{17B.1}$$

To get this, we have substituted for $\mathbf{b}_F(\theta) = \mathbf{T}\mathbf{b}_F$ from Eq. (17A.2), where $\mathbf{T} = \{\mathbf{I} + \theta(\mathbf{F}'\mathbf{F})^{-1}\}^{-1}$, isolated a quadratic term in $(\mathbf{b}_F - \boldsymbol{\beta}_F)$, taken the expectation $E(\mathbf{b}_F) = \boldsymbol{\beta}_F$ in the remaining terms, performed some cancellations, and regrouped. Application of the matrix results of Appendix 5A then leads to

$$\text{MSE}(\theta) = \sigma^2 \text{ trace } \{\mathbf{T}(\mathbf{F}'\mathbf{F})^{-1}\mathbf{T}'\} + \boldsymbol{\beta}_F'(\mathbf{T} - \mathbf{I})'(\mathbf{T} - \mathbf{I})\boldsymbol{\beta}_F. \qquad (17B.2)$$

The first term, being the sum of the diagonal terms of $V(\mathbf{T}\mathbf{b}_F) = \sigma^2\mathbf{T}(\mathbf{F}'\mathbf{F})^{-1}\mathbf{T}'$, is the sum of the variances of the elements of the ridge estimator $\mathbf{T}\mathbf{b}$. The second term is a ridge "squared bias" term. [Note that, if $\theta = 0$, then $\mathbf{T} = \mathbf{I}$ and the first term is the sum of the variances of the least squares estimators of the coefficients, while the second term vanishes. The value so attained would be MSE(0).] An existence theorem then states that there always exists a $\theta^* > 0$ such that $\text{MSE}(\theta^*) < \text{MSE}(0)$. The catch in this result is that its value depends on σ^2 and $\boldsymbol{\beta}_F$, neither of which is known. Thus although θ^* exists, there is no way of knowing whether or not we have attained a θ-value that provides lower MSE than MSE(0) in a specific practical problem.

APPENDIX 17C. CANONICAL FORM OF RIDGE REGRESSION

The canonical form of ridge regression requires a rotation of the $\boldsymbol{\beta}_F$-axes to new axes that are parallel to the principal axes of the sum of squares contours (See Figure 17C.1.) Let $\lambda_1, \lambda_2, \ldots, \lambda_r$ be the eigenvalues of $\mathbf{F}'\mathbf{F}$ and let

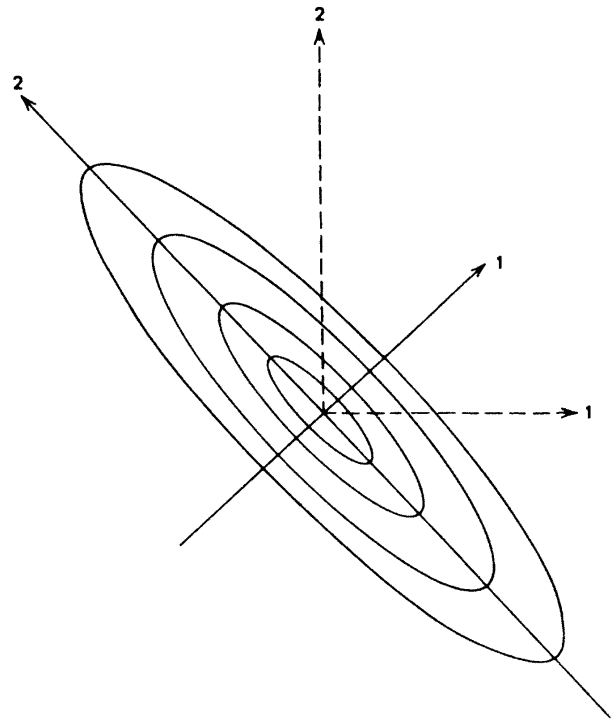

Figure 17C.1. The sum of squares surface contours are shown for $r = 2$: note that the center of the axial system has been moved to the point defined in the $\boldsymbol{\beta}_F$-space by the least squares estimates. The directions of the original axes are shown dashed, while the new rotated axes are solid.

$$\mathbf{G} = (\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_r) \tag{17C.1}$$

be the $r \times r$ matrix whose $r \times 1$ columns $\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_r$ are the eigenvectors of $\mathbf{F}'\mathbf{F}$. This means that the λ_i and $\mathbf{g}_i$ have the property that

$$\mathbf{F}'\mathbf{F}\mathbf{g}_i = \lambda_i \mathbf{g}_i, \qquad i = 1, 2, \ldots, r. \tag{17C.2}$$

To obtain these we solve the polynomial equation

$$\det(\mathbf{F}'\mathbf{F} - \lambda\mathbf{I}) = 0 \tag{17C.3}$$

for its r roots $\lambda_1, \lambda_2, \ldots, \lambda_r$. (In practical cases these roots are usually distinct. When they are not, which happens in some experimental design situations and occasionally in other cases, it means that certain sections of the sum of squares surface are circular or spherical, and no rotation is needed for axes already lying in that section.) We then substitute the λ_i in turn into the dependent Eqs. (17C.2) and choose the corresponding solutions $\mathbf{g}_i$ that are normalized so that $\mathbf{g}_i'\mathbf{g}_i = 1$. The following properties can be shown to hold:

$$\mathbf{G}'\mathbf{G} = \mathbf{G}\mathbf{G}' = \mathbf{I}, \tag{17C.4}$$

$$\mathbf{G}'\mathbf{F}'\mathbf{F}\mathbf{G} = \mathbf{\Lambda}, \tag{17C.5}$$

where $\mathbf{\Lambda}$ = diagonal $(\lambda_1, \lambda_2, \ldots, \lambda_r)$ is a diagonal matrix consisting of λ_i in the ith main diagonal position and zeros everywhere off the diagonal. Define $\mathbf{Q}$ and $\boldsymbol{\gamma}$ as follows:

$$\mathbf{Q} = \mathbf{F}\mathbf{G} \quad \text{and} \quad \boldsymbol{\gamma} = \mathbf{G}'\boldsymbol{\beta}_F, \tag{17C.6}$$

so that

$$\mathbf{F} = \mathbf{Q}\mathbf{G}' \quad \text{and} \quad \boldsymbol{\beta}_F = \mathbf{G}\boldsymbol{\gamma}. \tag{17C.7}$$

Then

$$\mathbf{F}'\mathbf{F} + \theta\mathbf{I} = \mathbf{G}(\mathbf{\Lambda} + \theta\mathbf{I})\mathbf{G}' \tag{17C.8}$$

and, because $\mathbf{G}^{-1} = \mathbf{G}'$, due to Eq. (17C.4),

$$(\mathbf{F}'\mathbf{F} + \theta\mathbf{I})^{-1} = \mathbf{G}(\mathbf{\Lambda} + \theta\mathbf{I})^{-1}\mathbf{G}'. \tag{17C.9}$$

The ridge estimator is thus

$$\mathbf{b}_F(\theta) = \mathbf{G}(\mathbf{\Lambda} + \theta\mathbf{I})^{-1}\mathbf{G}'\mathbf{F}'\mathbf{Y}, \tag{17C.10}$$

or, if we premultiply this equation by $\mathbf{G}'$,

$$\hat{\boldsymbol{\gamma}}(\theta) = (\mathbf{\Lambda} + \theta\mathbf{I})^{-1}\mathbf{Q}'\mathbf{Y}. \tag{17C.11}$$

If we write

$$c_i = (\mathbf{Q}'\mathbf{Y})_i \tag{17C.12}$$

for the ith component of the vector $\mathbf{Q}'\mathbf{Y}$ and invert the diagonal matrix $\mathbf{\Lambda} + \theta\mathbf{I}$, we obtain, for the components of $\hat{\boldsymbol{\gamma}}(\theta)$,

$$\hat{\gamma}_i(\theta) = c_i/(\lambda_i + \theta), \qquad i = 1, 2, \ldots, r. \tag{17C.13}$$

This is the *canonical form* of ridge regression. The ridge trace is usually plotted in terms of the β_F or β's, however. We can convert back to these via

$$\mathbf{b}_F(\theta) = \mathbf{G}\hat{\boldsymbol{\gamma}}(\theta), \tag{17C.14}$$

$$\mathbf{b}(\theta) = \{\text{Diagonal}(S_{jj}^{-1/2})\}\mathbf{G}\hat{\boldsymbol{\gamma}}(\theta). \tag{17C.15}$$

The canonical form is often seen in papers discussing ridge regression and provides a convenient computational form for making the ridge trace calculations.

Residual Sum of Squares

This can be written

$$\text{RSS}_\theta = \text{RSS}_0 + \theta^2 \sum_{i=1}^{r} \hat{\gamma}_i^2(\theta)/\lambda_i, \tag{17C.16}$$

where

$$\text{RSS}_0 = \left\{\sum Y_i^2 - n\overline{Y}^2\right\} - \sum_{i=1}^{r} c_i^2/\lambda_i \tag{17C.17}$$

is the minimum value attained by the least squares estimator $\hat{\boldsymbol{\gamma}}(0)$, which is given by $\theta = 0$. A possible way to select a value θ^* for θ is so that the difference $\text{RSS}_\theta - \text{RRS}_0$ is equal to some selected number or expression. One possibility is to set it equal to $s^2(r + 1)$. The rationale for this is that $\sigma^2(r + 1)$ is the expected value $E(\text{RRS}_\gamma - \text{RRS}_0)$ when the true value $\boldsymbol{\gamma}$ is used instead of $\hat{\boldsymbol{\gamma}}(\theta)$. This means that θ is selected in such a way that the squared distance represented by $\text{RRS}_\theta - \text{RSS}_0$ is the size we "expect" it to be. [Thus we attempt to keep the estimate on roughly the same confidence contour as that on which the true value falls on average. The direction from $\hat{\boldsymbol{\gamma}}(0)$ may, however, be different.] This provides

$$\theta^* = \left[s^2(r + 1) \Big/ \left\{\sum_{i=1}^{r} \gamma_i^2(\theta^*)/\lambda_i\right\}\right]^{1/2}, \tag{17C.18}$$

which could be solved for θ^*, perhaps iteratively in the way described around Eq. (17.3.4), provided convergence was achieved as in Eq. (17.3.5). Substitution of $\theta^* = 0$ on the right-hand side of Eq. (17C.18) provides the value from the first iteration, which might be good enough in some circumstances.

Mean Square Error

Equation (17B.2) becomes, in this notation

$$\text{MSE}(\theta) = \sum_{i=1}^{r} (\lambda_i\sigma^2 + \theta^2\gamma_i^2)/(\lambda_i + \theta)^2. \tag{17C.19}$$

Some Alternative Formulas

If we define

$$\hat{\boldsymbol{\gamma}}(0) = \boldsymbol{\Lambda}^{-1}\mathbf{Q}'\mathbf{Y} \tag{17C.20}$$

for the least squares estimate vector, and

$$\boldsymbol{\Delta} = (\boldsymbol{\Lambda} + \theta\mathbf{I})^{-1}\boldsymbol{\Lambda} \tag{17C.21}$$

for the diagonal matrix with elements $\delta_i = \lambda_i/(\lambda_i + \theta)$, we see that Eqs. (17.11), (17C.13), (17C.14), and (17C.19) become, respectively,

$$\hat{\boldsymbol{\gamma}}(\theta) = \boldsymbol{\Delta}\hat{\boldsymbol{\gamma}}(0), \tag{17C.22}$$

$$\hat{\gamma}_i(\theta) = \delta_i c_i/\lambda_i, \tag{17C.23}$$

$$\mathbf{b}_F(\theta) = \mathbf{G}\boldsymbol{\Delta}\hat{\boldsymbol{\gamma}}(0), \tag{17C.24}$$

$$\text{MSE}(\theta) = \text{trace}\{\sigma^2\boldsymbol{\Delta}^2\boldsymbol{\Lambda}^{-1} + (\mathbf{I} - \boldsymbol{\Delta})\boldsymbol{\gamma}\boldsymbol{\gamma}'(\mathbf{I} - \boldsymbol{\Delta})\}. \tag{17C.25}$$

These formulas exhibit the conversion from the least squares estimators to the ridge estimators.

EXERCISES FOR CHAPTER 17

A. Is ridge regression a least squares procedure?

B. Find the ridge regression solution for the data below for a general value of θ and for the straight line model $Y = \beta_0 + \beta_1 X + \epsilon$. Show that when θ is chosen as 0.4, the ridge solution fit is $\hat{Y} = 40 + 6.25Z = 40 + 1.5625X$, where Z is a centered and scaled form of X.

X	Y
−2	35
−1	40
−1	36
−1	38
0	40
1	43
2	45
2	43

C. Suggested readings for more on ridge regression (in addition to the basic Hoerl and Kennard papers mentioned in Section 17.1) are: D. W. Marquardt (1970), "Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation," *Technometrics*, **12,** 591–612; D. W. Marquardt and R. D. Snee (1975), "Ridge regression in practice," *The American Statistician,* **29,** 3–19; and "Why regression coefficients have the wrong sign," by G. M. Mullett, *Journal of Quality Technology,* **8,** 1976, 121–126.