Focus Article

Ridge regression

Gary C. McDonald*

Ridge regression is a popular parameter estimation method used to address the collinearity problem frequently arising in multiple linear regression. The formulation of the ridge methodology is reviewed and properties of the ridge estimates capsulated. In particular, four rationales leading to a regression estimator of the ridge form are summarized. Algebraic properties of the ridge regression coefficients are given, which elucidate the behavior of a ridge trace for small values of the ridge parameter (i.e., close to the least squares solution) and for large values of the ridge parameter. Further properties involving coefficient sign changes and rates-of-change, as functions of the ridge parameter, are given for specific correlation structures among the independent variables. These results help relate the visual behavior of a ridge trace to the underlying structure of the data.

© 2009 John Wiley & Sons, Inc. *WIREs Comp Stat* 2009 1 93–100

INTRODUCTION AND OVERVIEW

One of the most exciting research topics during the seventies and eighties was a newly proposed method, called ridge regression, for estimating multiple linear regression coefficients. Although there were earlier introductions of this methodology in the literature by Crone[1], Hoerl[2], and by Draper[3], its popularity rose dramatically with the publication of the article by Hoerl and Kennard[4] in *Technometrics*. Table 1 shows the number of articles related to ridge regression published in *Technometrics* (TECH), the *Journal of the American Statistical Association* (JASA), *Communications in Statistics—Theory and Methods* (CTM), and *Communication in Statistics–Simulation and Computation* (CSC) by decade. Two hundred and forty articles have been published, and about 39% of these were published in the eighties.

The popularity of this topic is attributable to the importance of the problem it addresses–collinearity in the multiple linear regression context–and the suitability of the methodology to be implemented easily in practice based on an examination of the so-called ridge trace, a plot of the estimated regression coefficients as a function of the ridge parameter. The ridge regression methodology, in fact, yields a class of biased estimators indexed by a scalar non-negative parameter. The challenge is to determine which estimator within this class to use in the context of a specific problem, i.e., determine a best choice for the ridge parameter. This quest to construct the best choice is one major impetus to the large number of research publications on this topic during the few decades following the seminal publications. Many of these quests involved large-scale simulation studies of stochastic estimators of the ridge parameter (i.e., estimators that are functions of the dependent variable), e.g., [5–9] to name just a few of the many. While many authors were quite enthusiastic about ridge regression, the methodology did spawn a significant number of critics, e.g., [10–13] among many others.

The negative impact of collinearity on the least squares (LS) estimator in a regression context is well known. Approaches to mitigate this impact were developed and many were centered on variable elimination, i.e., removing one or more of the independent variables so as to improve the conditioning of the resultant correlation matrix of the remaining independent variables. Ridge regression, on the other hand, provides a means of addressing the problem of collinearity without removing variables from the original set of independent variables. This proved to be a very attractive feature in some applications as shown by McDonald and Schwing[14] and Schwing and McDonald.[15] However, Hoerl and Kennard[16] did demonstrate how the ridge trace could be used in variable selection if so desired by the researcher. This ad hoc approach was used in the McDonald and Schwing[14] study relating air pollution indices to mortality rates and the subsequent problem-specific results compared quite favorably to that obtained by variable

*Correspondence to: mcdonald@oakland.edu

Department of Mathematics and Statistics, Oakland University, Rochester, MI 48309, USA

DOI: 10.1002/wics.014

selection using the Mallows' Cp statistic with all possible regressions. Gibbons and McDonald[17] show that overall estimates of the elasticity of air pollution on human mortality are higher using LS methods than with robust, bounded-influence, or ridge methods. The studies referenced in this paragraph provide very interesting applications of ridge regression analyses, construction and interpretation of ridge traces, and implementation of various strategies to choose ridge coefficients.

The purpose of this article is to provide a brief overview of the ridge regression method, the motivations for consideration of the ridge methodology, and a summary of algebraic properties of the ridge coefficients. Section on Formulation of the Ridge Regression Estimators provides the formulation of ridge regression estimators and the corresponding ridge trace. Section on Motivation for Ridge Regression provides four statistical model formulations or properties leading to solutions which, in form, are ridge regression estimators. Section on Algebraic Properties focuses on the algebraic properties of ridge regression coefficients and considers the deterministic properties which ensue from such representation. An understanding of these properties helps better understand the behavior of ridge coefficients, as functions of the ridge parameter, displayed in a ridge trace. Asymptotic (in the ridge parameter) properties of ridge coefficients are also noted. Concluding remarks comprises the last section of this article.

FORMULATION OF THE RIDGE REGRESSION ESTIMATORS

For the formulation of ridge regression, the description given by McDonald and Schwing[14] will be used. Multiple linear regression techniques have played a prominent role in the studies of associations between air pollution and mortality (and/or morbidity) rates, an important research area in which this author engaged in the early seventies. An LS approach may provide an adequate basis for overall prediction, but, when the explanatory variables are non-orthogonal, it frequently fails to give proper weight to the individual explanatory variables used as predictors. In many problems where data are not obtained from a well-designed or controlled experiment, as is the case in air pollution studies involving socioeconomic, weather, and other uncontrolled variables, non-orthogonality requires that estimation of individual effects be handled by techniques other than ordinary LS solutions. Reinke[18] pointed out these difficulties in air pollution models and suggested that ridge analysis provides a promising method for avoiding distortions as described shortly. The seminal papers of Hoerl and Kennard[4,16] and Marquardt[19] give an excellent description of the theory and applications of what has now been termed 'ridge regression'.

TABLE 1 | Number of Ridge Regression Articles

Journal	1970–1979	1980–1989	1990–1999	2000–2007	Total
TECH	27	15	4	7	53
JASA	6	9	4	5	24
CTM	22	56	26	26	130
CSC	1	14	13	5	33
Total	56	94	47	43	240

As has been shown by Hoerl and Kennard,[4] the estimates of regression coefficients tend to become too large in absolute values, and it is possible that some will even have the wrong sign. The chances of encountering such difficulties increase the more the prediction vectors deviate from orthogonality. Consider the standard model for multiple linear regression,

$$\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2\mathbf{I}_n$, and $\mathbf{x}$ is $(n \times p)$ and of full rank. Vectors and matrices will be denoted by bold symbols. The matrix $\mathbf{I}_n$ is the identity matrix with dimension n by n. The variables are assumed to be standardized so that $\mathbf{x}'\mathbf{x}$ is in the form of a correlation matrix, and the vector $\boldsymbol{\gamma} \equiv \mathbf{x}'\mathbf{y}$ is the vector of correlation coefficients of the response variable with each of the explanatory variables. The standardization is accomplished by subtracting the mean of the variable and then dividing by $(n-1)^{1/2}$ times the standard deviation of the variable, where n is the number of observations. Let

$$\widehat{\boldsymbol{\beta}} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y} \tag{2}$$

be the LS estimate of $\boldsymbol{\beta}$. The difficulties in this standard estimation are a direct consequence of the average distance between $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$. In particular, if L^2 is the squared distance between $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$, then the following hold:

$$\begin{aligned} L^2 &= (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}), \\ E(L^2) &= \sigma^2 \text{ trace } (\mathbf{x}'\mathbf{x})^{-1} \\ E(\widehat{\boldsymbol{\beta}}'\widehat{\boldsymbol{\beta}}) &= \boldsymbol{\beta}'\boldsymbol{\beta} + \sigma^2 \text{ trace } (\mathbf{x}'\mathbf{x})^{-1}. \end{aligned} \tag{3}$$

In terms of the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ of $\mathbf{x}'\mathbf{x}$, we can write

$$E(L^2) = \sigma^2 \sum_{i=1}^{p} \lambda_i^{-1} > \sigma^2 \lambda_p^{-1},$$
$$E(\widehat{\boldsymbol{\beta}}'\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}'\boldsymbol{\beta} + \sigma^2 \sum_{i=1}^{p} \lambda_i^{-1} > \boldsymbol{\beta}'\boldsymbol{\beta} + \sigma^2 \lambda_p^{-1}, \quad (4)$$

and when the error is normally distributed

$$\mathrm{V}ar(L^2) = 2\sigma^4 \sum_{i=1}^{p} \lambda_i^{-2} > 2\sigma^4 \lambda_p^{-2}. \quad (5)$$

As the vectors of $\mathbf{x}$ deviate further from orthogonality, λ_p becomes smaller, and $\widehat{\boldsymbol{\beta}}$ can be expected to be farther from the true parameter vector $\boldsymbol{\beta}$.

Ridge regression is an estimation procedure based upon

$$\widehat{\boldsymbol{\beta}}(k) = (\mathbf{x}'\mathbf{x} + k\mathbf{I})^{-1}\mathbf{x}'\mathbf{y},\ k \geq 0, \quad (6)$$

and has two aspects. The first is the ridge trace which is a two-dimensional plot of the $\widehat{\beta}_i(k)$ and the residual sum of squares (RSS), $\varphi(k)$, for values of k in the interval [0, 1]. While k can go beyond 1, it usually suffices to plot coefficients for $0 \leq k \leq 1$ to identify a value of k beyond which the ridge trace is quite stable. The trace serves to portray the complex interrelationships that exist between non-orthogonal prediction vectors and the effect of these interrelationships on the estimation of $\boldsymbol{\beta}$. The second aspect is the determination of a value of k that gives a better estimate of $\boldsymbol{\beta}$ by dampening the effect of (4). It should be noted that the estimators $\widehat{\boldsymbol{\beta}}(k)$ are biased when $k > 0$. Of course, at $k = 0$ these estimators reduce to those ordinary LS which are unbiased, i.e., $\widehat{\boldsymbol{\beta}}(0) = \widehat{\boldsymbol{\beta}}$.

The vector $\widehat{\boldsymbol{\beta}}(k)$ for $k > 0$ is shorter than $\widehat{\boldsymbol{\beta}}$, i.e., $[\widehat{\boldsymbol{\beta}}(k)]'[\widehat{\boldsymbol{\beta}}(k)] < \widehat{\boldsymbol{\beta}}'\widehat{\boldsymbol{\beta}}$. In fact, $[\widehat{\boldsymbol{\beta}}(k)]'[\widehat{\boldsymbol{\beta}}(k)]$ is a decreasing function in $k > 0$. For an estimate $\widehat{\boldsymbol{\beta}}(k)$ the RSS is given by

$$\begin{aligned} \varphi(k) &= [\mathbf{y} - \mathbf{x}\widehat{\boldsymbol{\beta}}(k)]'[\mathbf{y} - \mathbf{x}\widehat{\boldsymbol{\beta}}(k)] \\ &= \mathbf{y}'\mathbf{y} - [\widehat{\boldsymbol{\beta}}(k)]'\mathbf{x}'\mathbf{y} - k[\widehat{\boldsymbol{\beta}}(k)]'[\widehat{\boldsymbol{\beta}}(k)]. \quad (7) \end{aligned}$$

The $\mathbf{y}'\mathbf{y}$ term is the sum of squares of the dependent variable and is equal to 1 when the data are transformed as indicated in this section; the $[\widehat{\boldsymbol{\beta}}(k)]'\mathbf{x}'\mathbf{y}'$ is the sum of squares due to regression; and $k[\widehat{\boldsymbol{\beta}}(k)]'[\widehat{\boldsymbol{\beta}}(k)]$ is an adjustment term associated with the ridge analysis. The square of the correlation coefficient of the actual values, $\mathbf{y}$, and the predicted values, $\widehat{\mathbf{y}}$, is

$$\begin{aligned} R^2 &= (\widehat{\mathbf{y}}'\mathbf{y}\mathbf{y}'\widehat{\mathbf{y}})/(\widehat{\mathbf{y}}'\widehat{\mathbf{y}}) \\ &= [\widehat{\boldsymbol{\beta}}'(k)(\mathbf{x}'\mathbf{x})\widehat{\boldsymbol{\beta}}(k) + k\widehat{\boldsymbol{\beta}}'(k)\widehat{\boldsymbol{\beta}}(k)]^2 / \quad (8) \\ &\qquad [\widehat{\boldsymbol{\beta}}'(k)(\mathbf{x}'\mathbf{x})\widehat{\boldsymbol{\beta}}(k)], \end{aligned}$$

a decreasing function of k where $\mathbf{x}'\mathbf{x} = \mathbf{I}$, i.e., the explanatory variables are uncorrelated, then $\widehat{\boldsymbol{\beta}}(k) = (k+1)^{-1}\mathbf{x}'\mathbf{y} = (k+1)^{-1}\widehat{\boldsymbol{\beta}}$. In other words, the LS coefficients are uniformly scaled by the quantity $(k+1)^{-1}$. The relative values of the regression coefficients are then independent of the choice of k; i.e., $\widehat{\beta}_i(k)/\widehat{\beta}_j(k) = \widehat{\beta}_i/\widehat{\beta}_j, 1 \leq i, j \leq p, \widehat{\beta}_j \neq 0$, for all $k \geq 0$.

While not pursued in this article, many authors utilize the canonical decomposition approach to regression problems. This is very helpful when considering shrinkage and ridge techniques for parameter estimation as, for example, shown by Obenchain,[20] McDonald,[21] and others. Shrinkage and ridge techniques are really both shrinkage methods in canonical variables. The canonical approach provides insight into the nature of ridge regression and leads to meaningful generalizations. One such generalization is a two-parameter family, introduced by Obenchain, of the form

$$\widehat{\boldsymbol{\beta}}(k, q) = \left[\mathbf{x}'\mathbf{x} + k(\mathbf{x}'\mathbf{x})^q\right]^{-1}\mathbf{x}'\mathbf{y}. \quad (9)$$

Goldstein and Smith,[22] Crone,[1] and Mayer and Willke[23] also consider estimators from this two-parameter family. Of course, $\widehat{\boldsymbol{\beta}}(k, 0)$ reduces to $\widehat{\boldsymbol{\beta}}(k)$ of (6).

MOTIVATION FOR RIDGE REGRESSION

Several properties of the ridge estimator justify its consideration as an alternative to the LS estimator. Four such properties are noted in this section.

Hoerl and Kennard[4] show that there exists a range of k values, say $0 < k < k^*$, for which the total mean squared error for the ridge estimator is smaller than the corresponding LS quantity. Theobald[24] extends this result to a more general loss function. Analogous results for the mean squared error of prediction and mean squared error for each coefficient evolve as special cases of Theobald's extension. These derivations demonstrate existence of such a k^*. However, the proofs are not constructive, e.g., they do not indicate how to determine k^*.

© 2009 John Wiley & Sons, Inc.

Lindley and Smith[25] show that if $\mathbf{y} \sim N(\mathbf{x}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ and $\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma_\beta^2\mathbf{I})$, then $\widehat{\boldsymbol{\beta}}(k)$ is the Bayes estimator where $k = \sigma^2/\sigma_\beta^2$. The vector $\widehat{\boldsymbol{\beta}}(k)$ is the mean vector of the posterior distribution of $\boldsymbol{\beta}$. As $\sigma_\beta^2 \to \infty$, then $k \to 0$ and the Bayes estimator $\widehat{\boldsymbol{\beta}}(k)$ approaches the LS estimator $\widehat{\boldsymbol{\beta}}(0)$. As $\sigma_\beta^2 \to 0$, then $k \to \infty$ and the Bayes estimator $\widehat{\boldsymbol{\beta}}(k)$ approaches the zero vector.

The estimator $\widehat{\boldsymbol{\beta}}(k)$ is a constrained LS estimator; i.e., $\widehat{\boldsymbol{\beta}}(k)$ minimizes the RSS subject to a constraint on the length of the estimated coefficient vector. This follows from a more general result. Assume the linear regression model (1) and $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \mathbf{U}$, a known positive definite covariance matrix. An estimator of $\boldsymbol{\beta}$ which minimizes the generalized sum of squares

$$f(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{x}\boldsymbol{\beta})'\mathbf{U}^{-1}(\mathbf{y} - \mathbf{x}\boldsymbol{\beta}) \qquad (10)$$

subject to a quadratic constraint expressed as

$$h(\boldsymbol{\beta}) \geq 0, \qquad (11)$$

where $h(\boldsymbol{\beta}) \equiv r - \boldsymbol{\beta}'\mathbf{A}\boldsymbol{\beta}$, $\mathbf{A}$ a known $p \times p$ symmetric semidefinite matrix, and r a fixed non-negative scalar, is shown by McDonald[26] with suitable conditions to be

$$\widehat{\boldsymbol{\beta}}(k) = (\mathbf{x}'\mathbf{U}^{-1}\mathbf{x} + k\mathbf{A})^{-1}\mathbf{x}'\mathbf{U}^{-1}\mathbf{y}, \qquad (12)$$

where $g[\widehat{\boldsymbol{\beta}}(k)] \equiv [\widehat{\boldsymbol{\beta}}(k)]'\mathbf{A}[\widehat{\boldsymbol{\beta}}(k)] = r$. Choosing $\mathbf{U} = \mathbf{I}_n$ and $\mathbf{A} = \mathbf{I}_p$ reduces $\widehat{\boldsymbol{\beta}}(k)$ of (12) to that of (6).

The least absolute shrinkage and selection operator (lasso), proposed by Tibshirani,[27] is a shrinkage method somewhat like the ridge method. The parameter estimates are chosen to minimize the RSS subject to $\sum_{i=1}^{p} |\beta_i| \leq s$. The L_2 ridge penalty is replaced by the L_1 lasso penalty. Hastie et al.,[28] discuss and compare lasso, ridge, partial least squares (PLS), principal components regression (PCR), and subset regression. They find that PLS, PCR, and ridge regression tend to behave similarly. They note that ridge regression may be preferred because it shrinks smoothly rather than in discrete steps.

Leamer and Chamberlain[29] showed that the ridge regression estimator is expressible as a weighted average of 2^p constrained LS estimators, or equivalently, as a weighted average of the LS estimators derived from all possible subset regressions. Gibbons and McDonald[30] show this approach to ridge regression leads to the representation of ridge regression estimators as a ratio of polynomials in k. This representation provides an explicit understanding of the initial rate-of-change of the ridge regression coefficients and the asymptotic behavior of the ridge trace. In particular, the behavior of the ridge trace near $k = 0$ is governed by the relationship between the unconstrained LS estimator and the p LS estimators with exactly one constraint. Hence, for any application, insight into the instability of the ridge trace can be obtained by evaluating $p + 1$ regression equations. The asymptotic behavior of the ridge trace is governed by the marginal correlations of the explanatory variables with the dependent variable (i.e., the p constrained LS estimators with exactly $p - 1$ constraints).

This approach to ridge regression also shows that the ridge trace provides a mechanism for examining the sensitivity of regression estimates to the choice of variables in the model. The ridge trace spans all possible subset regressions, and the weights applied to the subset regression estimators (constrained LS estimators) vary systematically with the ridge parameter k. The weights shift from LS estimators with no or few constraints to LS estimators with many constraints as k increases. The properties of ridge regression estimators–namely, increased bias and decreased variance–are properties shared by the constrained LS estimators.[31]

The Bayes formulation of the multiple linear regression model provides a framework for understanding the weights assigned to the subset regressions. The weights are shown to incorporate sample information and prior information in an intuitively appealing way. A distinct pattern to the weights was observed and analytically verified in the Gibbons and McDonald[30] article.

These properties, however, do not define explicitly an appropriate k value to use in a specific application. The range of k values for which the ridge estimator dominates the LS estimator in mean squared error depends on the unknown $\boldsymbol{\beta}$ as well as σ^2. The Bayes interpretation of ridge regression yields a k value that is a ratio of unknown variances. Also, the constrained LS approach does not define a specific k value in practice because an explicit constraint on the length of the estimated coefficient vector is unknown in most applications.

Choice of a ridge parameter to use in a particular application is often based on a visual analysis of the ridge trace with an assessment of goodness-of-fit statistics as a function of k. The analysis of the ridge trace involves identifying the smallest reasonable value of k at which the trace depicts somewhat stable behavior, i.e., behavior that is more like that expected from an orthogonal design. For example, rapid changes in individual coefficients and criss-crossing of coefficients have mitigated, and coefficients have assumed intuitively appealing signs. The squared length of the coefficient vector has flattened out and the RSS and R^2 have not deviated excessively from the corresponding LS values. The resultant ridge regression estimator of β is interior to a suitably specified confidence ellipsoid.

Marquardt[19] has recommended use of the variance inflation factors (VIFs) as a means of narrowing the search of ridge estimators. This criteria have proved useful with many applications. The parameter VIFs for the ridge estimator are the diagonal elements of $(\mathbf{x}'\mathbf{x} + k\mathbf{I})^{-1}\,(\mathbf{x}'\mathbf{x})\,(\mathbf{x}'\mathbf{x} + k\mathbf{I})^{-1}$. A rule-of-thumb suggested by Marquardt for choosing the amount of bias to allow with ill-conditioned data is that the maximum VIF usually should be between 1 and 10. However, this criterion only depends on the $\mathbf{x}'\mathbf{x}$ matrix and not on the dependent variable, $\mathbf{y}$. As such, for any specific choice of k, it is always possible to construct a set of true regression coefficients for which this approach would fail to improve the mean squared error. Strawderman[32] developed a stochastic estimator of k which leads to an estimator which does dominate LS.

There are a large number of stochastic (i.e., dependent on the vector $\mathbf{y}$) analytical approaches to the choice of k. Theoretical properties of such estimators are, in general, not obtainable. So their properties are most often inferred by computer simulation. Many such studies are available in the literature:[5,33,34,35,36], and many others. As noted by McDonald and Galarneau[33], the performance of the evaluated ridge-type estimators, as well as the potential performance of any ridge-type estimator, depends on three critical factors: the variance of the random error, the correlations among the explanatory variables, and the unknown coefficient vector. Gibbons[5] provides a detailed summary of the performance of twelve ridge-type estimators as these critical factors are systematically varied.

ALGEBRAIC PROPERTIES

Section on Motivation for Ridge Regression deals with four rationales that lead to an estimator of β that assumes the form of the ridge regression estimator. As noted earlier, Leamer and Chamberlain[29] show the ridge estimator belongs to the class of search estimators and, in particular, each ridge coefficient is a rational function of k.

From these results, characteristics of ridge regression coefficients that hold in general can be deduced. For example, for large k, the ridge coefficient vector will tend to be a shrunken multiple of the marginal correlation vector. Additionally, the behavior of the ridge trace near $k = 0$, the LS point, is determined in a well-defined manner by the constrained regression estimators with one variable at a time constrained to be zero. See Gibbons and McDonald[30] for explicit examples. Characterizing ridge trace behavior for the 'in between' k-values requires additional assumptions on the data structure which is the topic of this section.

Following the rational function representation, McDonald[37] shows the ith ridge coefficient can be expressed as

$$\widehat{\beta}_i(k) = P_i^{(p-1)}(k)/Q^{(p)}(k), \qquad i = 1, \ldots, p, \qquad (13)$$

where $Q^{(p)}(k) \equiv \Pi_{i=1}^{p}(\lambda_i + k)$ is a monic polynomial in k of degree exactly p, and $P_i^{(p-1)}(k)$ is a polynomial in k of degree at most $p - 1$. The highest order term of $P_i^{(p-1)}(k)$ is $\gamma_i k^{p-1}$ and the coefficients for terms involving lower powers of k can be expressed as linear combinations of restricted LS estimates.

Algebraic properties of ridge coefficients which follow from the rational function characterization include:

(**a**) Once the eigenvalues of $\mathbf{x}'\mathbf{x}$ have been determined the ridge estimator $\widehat{\beta}(k)$ can be determined uniquely, as a function of k, by evaluating (6) at p distinct values of k.

(**b**) The exact number of sign changes experienced by an individual $\widehat{\beta}_i(k)$, $k \geq 0$, is no greater than the number of distinct positive real roots of $P_i^{(p-1)}(k)$.

(**c**) Asymptotically in k, $\widehat{\beta}_i(k) \sim \gamma_i/k$, and so $\widehat{\beta}_i(k)$ will eventually assume the sign of γ_i. Marquardt[19] has noted that the angle between $\widehat{\beta}(k)$ and γ is a continuous monotone decreasing function of k which converges to zero as $k \to \infty$.

(**d**) If $\gamma_i < \gamma_j$ and k is sufficiently large, then $\triangle_i(k) > \triangle_j(k)$, where

$$\triangle_i(k) \equiv \partial\widehat{\beta}_i(k)/\partial k, \; i = 1, 2, \ldots, p. \qquad (14)$$

Additional properties of ridge coefficients may hold if the $\mathbf{x}'\mathbf{x}$ matrix has a specific form. McDonald[37] considers the case where $\mathbf{x}'\mathbf{x}$ has intraclass correlation structure, i.e., $\mathbf{x}'\mathbf{x} = (x_{ij})$, where $x_{ii} = 1$, $x_{ij} = \rho$ for $i \neq j$, $1 \leq i, j \leq p$ and $-(p-1)^{-1} < \rho < 1$. In this case the eigenvectors do not depend upon ρ. Any p mutually orthogonal vectors of unit length, the first of which has equal components, suffices. All problems with two explanatory variables, i.e., $p = 2$, are of this form. In addition, this structure may provide an adequate approximation for more complex problems when $p \geq 3$. The ridge estimators now take the form

$$\begin{aligned}\widehat{\beta}(k) &= (\mathbf{x}'\mathbf{x} + k\mathbf{I})^{-1}\mathbf{x}'\mathbf{y} = (1 + k - \rho)^{-1} \\ &\quad \times [\mathbf{I} - \rho\{1 + k + (p-1)\rho\}^{-1}\mathbf{J}]\gamma, \ k \geq 0,\end{aligned} \tag{15}$$

where $\mathbf{I}$ is the $p \times \rho$ identity matrix, $\mathbf{J}$ is the $p \times \rho$ unity matrix, and $\gamma' = (\gamma_1, \ldots \gamma_p)$ with γ_i being the correlation between $\mathbf{x}_i$ and $\mathbf{y}$. Letting $\bar{\gamma} = (\Sigma\gamma_i)/p$, summing over i from one to p, the individual ridge coefficients can be written as

$$\widehat{\beta}_i(k) = (1 + k - \rho)^{-1}\gamma_i - C, \quad i = 1, 2, \ldots, p, \tag{16}$$

where

$$\begin{aligned}C &= C(k, \rho, \gamma) = \rho p \bar{\gamma}(1 + k - \rho)^{-1} \\ &\quad \times \{(1 + k - \rho) + p\rho\}^{-1}.\end{aligned} \tag{17}$$

Several interesting properties follow directly from this representation. For all i and j , $1 \leq i, j \leq p$, and $0 \leq k < \infty$,

(a) $\widehat{\beta}_i(k) < \widehat{\beta}_j(k)$ iff $\gamma_i < \gamma_j$;

(b) $\triangle_i(k) > \triangle_j(k)$ iff $\gamma_i < \gamma_j$.

(c) If $\gamma_i \widehat{\beta}_i(0) > 0$, then $\widehat{\beta}_i(k)$ has the same sign as $\widehat{\beta}_i(0)$ for $0 \leq k < \infty$.

(d) If $\gamma_i \widehat{\beta}_i(0) < 0$, then there exists a $k' > 0$, such that the sign of $\widehat{\beta}_i(k)$ is reversed at k' and only k', i.e. if $\widehat{\beta}_i(0) > 0$, then $\widehat{\beta}_i(k) > 0$ for $k < k'$ and $\widehat{\beta}_i(k) \leq 0$ for $k \geq k'$, and similarly if $\widehat{\beta}_i(0) < 0$.

These properties have direct interpretations with respect to a ridge trace. Part (a) states that there is no criss-crossing of the paths of the individual coefficient estimates (unless, of course, two or more coincide for all k). Part (b) explicitly relates the 'instability' (or rate-of-change) of a ridge trace with the vector of correlation coefficients between the explanatory variables and dependent variable. According to (c) and (d), a sign change in a coefficient estimate occurs only if the LS estimate differs in sign from the correlation coefficient of the corresponding explanatory variable and dependent variable. In this sense, if a sign change occurs, it does change in the appropriate direction.

Gibbons[38] derives properties for $p = 3$ and $p = 4$ when $\mathbf{x}'\mathbf{x}$ has Toeplitz correlation structure, i.e., $\mathbf{x}'\mathbf{x} = (x_{ij})$, where $x_{ij} = \rho^{|i-j|}$, $|\rho| < 1$.

When $p = 3$,

(a) The first and third ridge coefficients are ordered in the same way as the correlations of the first and third explanatory variables with $\mathbf{y}$ for all k. Hence there is no criss-crossing of the first and third ridge regression estimates.

(b) The rate-of-change of the ridge coefficients for the first and third variables is related to their marginal correlations with $\mathbf{y}$.

(c) The first and third ridge coefficients behave as if the design matrix has intraclass correlation structure.

(d) The first and second (or second and third) ridge coefficients are not necessarily ordered by the marginal correlations with $\mathbf{y}$.

(e) The maximum number of sign changes in an individual coefficient is two.

When $p = 4$,

(a) No ordered relationships between the ridge coefficients and the marginal correlations with $\mathbf{y}$ hold in general.

(b) The first and fourth ridge coefficients are ordered by their marginal correlations if the marginal correlations of the second and third variables with $\mathbf{y}$ are equal.

(c) Similarly, the second and third ridge coefficients are ordered by their marginal correlations if the marginal correlations of the first and fourth variables with $\mathbf{y}$ are equal.

Zhang and McDonald[39] extend this line of inquiry for the general Toeplitz correlation structure ($p > 4$). There appear to be no ordered relationships between the ridge coefficients and the marginal correlations, γ_i, that hold in general. However, the ridge coefficients $\widehat{\beta}_i(k)$ and $\widehat{\beta}_j(k)$ are ordered when $i + j = p + 1$ by their marginal correlations if the marginal correlations of all other pairs of γ_{i*} and γ_{j*} are equal, where $i^* + j^* = p + 1$.

They also consider the case where $\mathbf{x}'\mathbf{x}$ represent a correlation matrix which has banded correlation structure with bandwidth 3 (tri-diagonal matrix). The tri-diagonal matrix is a special case of the symmetric Toeplitz matrix in which all elements are zero except those on the super, principal, and sub-diagonal. For their analysis, $x_{ii} = 1$, $x_{ij} = \rho$ for $|i - j| = 1$. It can be shown that

(a) $\widehat{\beta}_1(k) < \widehat{\beta}_3(k)$ when $\gamma_1 < \gamma_3$, and $k > \sqrt{2} - 1$; or when $\gamma_1 < \gamma_3, k < \sqrt{2} - 1$, and $-(k+1)/\sqrt{2} < \rho < (k+1)/\sqrt{2}$,

(b) $\triangle_1(k) > \triangle_3(k)$ iff $\gamma_1 < \gamma_3$.

Hence the order of the rates-of-change on the ridge trace for the first and third coefficients are exactly opposite to the correlations between the explanatory variables and the response variable. There is no analogous relationship between $\widehat{\beta}_1(k)$ and $\widehat{\beta}_2(k)$ (or between $\widehat{\beta}_2(k)$ and $\widehat{\beta}_3(k)$).

CONCLUDING REMARKS

As indicated in Table 1, there has been a tremendous amount of research devoted to the topic of ridge regression. It is a methodology that should be a standard tool for applied statisticians. Indeed, the methodology is making its way into many of the methods textbooks (e.g., see [40,41] and others). Ridge regression, and the associated ridge trace, provides the analyst with a systematic way to scan a large class of biased estimators. By so doing, it frequently leads the analyst to meaningful estimates of regression coefficients that can be used in practice for assessing changes of the dependent variable with respect to the independent variables and also predict values of the response at specified points in the design space.

Much of the literature on ridge regression is devoted to identifying an optimal ridge parameter, k, to be used in practice–if not optimal, at least a k-value that will insure the ridge estimator has lower mean squared error than the LS estimator. Although there are many good methods for choosing an 'optimal' k-value, no uniform winner has emerged. A more pragmatic goal is to identify ridge estimates that are more useful than the LS estimates. Such a goal would include tradeoffs in bias with meaningful magnitudes and signs of regression coefficients; tradeoffs between bias and variance of estimates; tradeoffs between minimizing the RSS vs. attaining a reasonably small RSS; tracking ridge estimators within confidence ellipsoids for the regression parameters; etc.

In summary, ridge regression has strong bases for consideration as shown in section on Motivation for Ridge Regression. There are many scientific approaches on the proper choice of the ridge parameter. The ultimate choice of k for a specific application involving collinear explanatory variables still remains part art and part science. For over three decades ridge regression has proved to be a valuable tool for use by applied statisticians and should be routinely explored in a collinear multiple regression context.

REFERENCES

1. Crone L. The singular value decomposition of matrices and cheap numerical filtering of systems of linear equations. *J Franklin Inst* 1972, 294:133–136.
2. Hoerl AE. Application of ridge analysis to regression problems. *Chem Eng Prog* 1962, 58:54–59.
3. Draper NR. "Ridge analysis" of response surfaces. *Technometrics* 1963, 5:469–479.
4. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 1970a, 12:55–67.
5. Gibbons DI. A simulation study of some ridge estimators. *J Am Stat Assoc* 1981, 76:131–139.
6. Hoerl AE, Kennard RW, Baldwin KF. Ridge regression: some simulations. *Commun Stat Theory Methods* 1975, 4:105–123.
7. Lawless JF, Wang P. A simulation study of ridge and other regression estimators. *Commun Stat Theory Methods* 1976, 5:307–323.
8. Wichern DW, Churchill GA. A comparison of ridge estimators. *Technometrics* 1978, 20:301–312.
9. Zhang J, Ibrahim M. A simulation study on SPSS ridge regression and ordinary least squares regression procedures for multicollinearity data. *J Appl Stat* 2005, 32:571–588.
10. Conniffe D, Stone J. A critical view of ridge regression. *Statistician*, 1974, 22:181–187.
11. Draper NR, Van Nostrand RC. Ridge regression and James-Stein estimation: review and comments. *Technometrics* 1979, 21:451–466.
12. Rozeboom WW. Ridge regression: bonanza or beguilement? *Psychol Bull* 1979, 8:242–249.
13. Smith G, Campbell F. A critique of some ridge regression methods. *J Am Stat Assoc* 1980, 75:74–103.
14. McDonald GC, Schwing RC. Instabilities of regression estimates relating air pollution to mortality. *Technometrics* 1973, 15:463–481.
15. Schwing RC, McDonald GC. Measures of association of some air pollutants, natural ionizing radiation and cigarette smoking with mortality rates. *Sci Total Environ* 1976, 5:139–169.
16. Hoerl AE, Kennard RW. Ridge regression: applications to nonorthogonal problems. *Technometrics* 1970b, 12:69–82.

17. Gibbons DI, McDonald GC. Illustrating regression diagnostics with an air pollution and mortality model. *Comput Stat Data Anal*, 1983, 1:201–220.

18. Reinke WA. Multivariate and dynamic air pollution models. *Arch Environ Health* 1969, 18:481–484.

19. Marquardt DW. Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation. *Technometrics* 1970, 12:591–612.

20. Obenchain RL. Ridge analysis following a preliminary test of the shrunken hypothesis. *Technometrics* 1975, 17:431–441.

21. McDonald GC. Discussion of: ridge analysis following a preliminary test of the shrunken hypothesis. *Technometrics* 1975, 17:443–445.

22. Goldstein M, Smith AFM. Ridge-type estimators for regression analysis. *J Roy Stat Soc B* 1974, 36:284–291.

23. Mayer LS, Willke TA. On biased estimation in linear models. *Technometrics* 1975, 15:497–508.

24. Theobald CM. Generalization of mean square error applied to ridge regression. *J Roy Stat Soc B* 1974, 36:103–106.

25. Lindley DV, Smith AFM. Bayes estimates for the linearestimation and nonlinear estimation. *J Roy Stat Soc B* 1972, 34:1–18.

26. McDonald GC. Ridge estimators as constrained generalized least squares. In: Gupta ESS, Berger JO eds. *Statistical Decision Theory and Related Topics III*, Vol. 2, New York, NY: Academic Press; 1982, 183–191.

27. Tibshirani R. Regression shrinkage and selection via the lasso. *J Roy Stat Soc B* 1996, 58:267–288.

28. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer-Verlag, Inc.; 2001.

29. Leamer EE, Chamberlain G. A Bayesian interpretation of pretesting. *J Roy Stat Soc B* 1976, 38:85–94.

30. Gibbons DI, McDonald GC. A rational interpretation of the ridge trace. *Technometrics* 1984, 26:339–346.

31. Hocking RR. The analysis and selection of variables in linear regression. *Biometrics* 1976, 32:1–49.

32. Strawderman WE. Minimax adaptive generalized ridge regression estimators. *J Am Stat Assoc* 1978, 73:623–627.

33. McDonald GC, Galarneau DI. A Monte Carlo evaluation of some ridge-type estimators. *J Am Stat Assoc* 1975, 70:407–416.

34. Trenkler D, Trenkler G. A simulation study comparing some biased estimators in the linear model. *Comput Stat Q* 1984, 1:45–60.

35. Clark AE, Troskia CG. Ridge regression -a simulation study. *Commun Stat Simulat Comput*, 2006, 35:605–619.

36. Alkhamisi MA, Shukur G. A Monte Carlo study of recent ride parameters. *Commun Stat Simulat Comput* 2007, 36:535–547.

37. McDonald GC. Some algebraic properties of ridge coefficients. *J Roy Stat Soc B* 1980, 42:31–34.

38. Gibbons DI. Some characterizations of the ridge trace. *Commun Stat Theory Methods* 1984, 13:173–182.

39. Zhang R, McDonald GC. Characterization of ridge trace behavior. *Commun StatTheory Methods* 2005, 34:1487–1501.

40. Mendenhall W, Sincich T. *A Second Course in Statistics: Regression Analysis*, 6th ed.. New Jersey: Pearson Education, Inc.; 2003.

41. Chatterjee S, Hadi AS. *Regression Analysis by Example*. 4th ed. New York, NY: John Wiley & Sons, Inc. 2006.