International Statistical Review (2008), 76, 1, 89–105 doi:10.1111/j.1751-5823.2007.00041.x

Anomalies in the Foundations of Ridge Regression

Donald R. Jensen and Donald E. Ramirez

Department of Mathematics, University of Virginia, Charlottesville, Virginia 22904–4137, USA
E-mails: djensen@vt.edu, der@virginia.edu

Summary

Errors persist in ridge regression, its foundations, and its usage, as set forth in Hoerl & Kennard (1970) and elsewhere. Ridge estimators need not be minimizing, nor a prospective ridge parameter be admissible. Conventional estimators are not LaGrange's solutions constrained to fixed lengths, as claimed, since such solutions are singular. Of a massive literature on estimation, prediction, cross–validation, choice of ridge parameter, and related issues, little emanates from constrained optimization to include inequality constraints. The problem traces to a misapplication of LaGrange's Principle, unrecognized singularities, and misplaced links between constraints and ridge parameters. Alternative principles, based on condition numbers, are seen to validate both conventional ridge and *surrogate ridge* regression to be defined. Numerical studies illustrate that ridge regression as practiced often exhibits pathologies it is intended to redress.

Key words: Constrained optimization; incomplete use of LaGrange's method; non-singular distributions; alternative foundations.

1 Introduction

For a model $\boldsymbol{Y} = \boldsymbol{X}\beta + \epsilon$ with zero-mean, homoscedastic, and uncorrelated errors, the ordinary least squares (*OLS*) solutions $\widehat{\boldsymbol{\beta}}_L$ solve the p equations $\{\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}\}$ minimizing $Q(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$. In ill-conditioned models, $\widehat{\boldsymbol{\beta}}_L$ may exhibit excessive length, inflated variances, instability, and other intrinsic deficiencies, prompting Hoerl (1962, 1964) to consider $\widehat{\boldsymbol{\beta}}_R = \{\widehat{\boldsymbol{\beta}}_{Rk} = (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I}_p)^{-1}\boldsymbol{X}'\boldsymbol{Y}; k \geq 0\}$, with *ridge parameter k*, for use in chemical engineering. These are labeled *ridge regression* in statistics, despite earlier usage in numerical analysis (Levenberg, 1944; Riley, 1955). In view of excessive lengths of $\widehat{\boldsymbol{\beta}}_L$, and that *OLS* "does not have built into it a method for portraying sensitivity of the solutions to the estimation criterion," Hoerl & Kennard (1970) sought foundations for ridge regression beyond limitations intrinsic to Gauss's Principle and *OLS*. They assert (i) that $\widehat{\boldsymbol{\beta}}_{Rk}$ minimizes $Q(\boldsymbol{\beta})$ constrained to $\{\boldsymbol{\beta}'\boldsymbol{\beta} = c^2\}$; (ii) that constrained solutions are linear in $\widehat{\boldsymbol{\beta}}_L$; (iii) that the analysis may proceed on choosing $k = \ell$ and $\widehat{\boldsymbol{\beta}}_{R\ell}$ as the solution; and (iv) that the implied constraint may be recovered as $\{c_\ell^2 = \widehat{\boldsymbol{\beta}}'_{R\ell}\widehat{\boldsymbol{\beta}}_{R\ell}\}$. Confusion nonetheless persists. Bunke (1975), Hocking (1976), Tibshirani (1996), and others claim the constraints $\{\boldsymbol{\beta}'\boldsymbol{\beta} \leq c^2\}$ adopted by Balakrishnan (1963) in control theory, despite the disclaimer of Hoerl & Kennard (1970, p. 64). These topics are addressed subsequently.

Ridge estimators abound, based on estimative, predictive, cross–validative, graphic, and other criteria, offering widely disparate choices for k and $\widehat{\boldsymbol{\beta}}_{Rk}$ to be noted. A vast literature, spanning four decades, rests ostensibly on constrained optimization, as claimed in Hoerl & Kennard (1970). Unfortunately, these premises are false. Our purposes here are to review essential literature; to document critical errors; to disenfranchise ridge regression from constrained optimization; to reveal its intrinsically *ad hoc* character; and to advance other approaches having attractive properties. At issue are misuse of LaGrange's Principle; unrecognized singularities; and misplaced links between constraints and ridge parameters, mistaken to be one-to-one. To the contrary, ridge solutions $\widehat{\boldsymbol{\beta}}_{Rk}$ need not be minimizing, whereas admissible values for k, thus $\widehat{\boldsymbol{\beta}}_{Rk}$, must be sought. These misconceptions pervade much of the literature and are explicit in Marquardt (1970), Marquardt & Snee (1975), Golub *et al.* (1979), and Nostrand (1980), for example. In short, the collective body of ridge regression rests on little more than heuristics, despite its pervasive use in practice. Accordingly, we seek to supplant the discredited foundations using conditioning of linear systems. An outline follows.

Supporting materials comprise Section 2, to include notation, basic matrices and vectors, and a concise canonical form. Section 3 reviews essentials of ridge regression; it retraces critical anomalies and their consequences; and it re-examines an ill-conditioned data set treated elsewhere. Section 4 seeks rationale alternative to constrained optimization, specifically, the conditioning of linear systems. These approaches then are compared and illustrated through a continuing case study. Section 5 offers a summary and conclusions. A survey of LaGrange optimization is relegated to an Appendix.

2 Preliminaries

2.1 Notation

Vectors and matrices are set in bold type; $\mathbb{R}^p$ designates Euclidean p-space; the Euclidean norm of $\boldsymbol{x} \in \mathbb{R}^p$ is $\|\boldsymbol{x}\|$; and the transpose, inverse, trace, and determinant of $\boldsymbol{A}$ are $\boldsymbol{A}'$, $\boldsymbol{A}^{-1}$, $\mathrm{tr}(\boldsymbol{A})$, and $|\boldsymbol{A}|$. Groups of note include $\mathcal{O}(p)$ as the real orthogonal group; special arrays are the $(p \times p)$ identity $\boldsymbol{I}_p$ and the diagonal matrix $\boldsymbol{D}_a = \mathrm{Diag}(a_1, \ldots, a_p)$. Given $\boldsymbol{X}(n \times p)$ of rank $p \le n$, its *singular decomposition* is $\boldsymbol{X} = \boldsymbol{P}\boldsymbol{D}_\xi\boldsymbol{Q}' = \sum_{i=1}^{p} \xi_i \boldsymbol{p}_i \boldsymbol{q}_i'$, where $\boldsymbol{D}_\xi = \mathrm{Diag}(\xi_1, \ldots, \xi_p)$, $\boldsymbol{P} = [\boldsymbol{p}_1, \ldots, \boldsymbol{p}_p]$ is semi-orthogonal with $\boldsymbol{P}'\boldsymbol{P} = \boldsymbol{I}_p$, and $\boldsymbol{Q} \in \mathcal{O}(p)$. Here $\{\xi_1 \ge \cdots \ge \xi_p > 0\}$ are ordered *singular values* of $\boldsymbol{X}$, and columns of $\boldsymbol{P} = [\boldsymbol{p}_1, \ldots, \boldsymbol{p}_p]$ and $\boldsymbol{Q} = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_p]$ comprise its *left-* and *right-singular vectors*. If $\boldsymbol{Y} \in \mathbb{R}^n$ is random, designate its mean vector, dispersion and correlation matrices as $\mathrm{E}(\boldsymbol{Y})$, $\mathrm{V}(\boldsymbol{Y})$, and $C(\boldsymbol{Y})$, and its law of distribution as $\mathcal{L}(\boldsymbol{Y})$.

The Solutions: The following glossary identifies continuing notation:

- $\widehat{\boldsymbol{\beta}}_L = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$: The *OLS* estimators;
- $\widehat{\boldsymbol{\beta}}_{Rk} = \widehat{\boldsymbol{\beta}}_R(k) = (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I}_p)^{-1}\boldsymbol{X}'\boldsymbol{Y}$: Solutions of Hoerl (1962, 1964);
- $\widehat{\boldsymbol{\beta}}_c$: Solutions satisfying $\{\widehat{\boldsymbol{\beta}}_c'\widehat{\boldsymbol{\beta}}_c = c^2\}$;
- $\widehat{\boldsymbol{\beta}}_c^o$: Solutions satisfying $\{\widehat{\boldsymbol{\beta}}_c^{o\prime}\widehat{\boldsymbol{\beta}}_c^o \le c^2\}$.

This notation is intended (i) to distinguish genuinely different solutions, and (ii) to enable unambiguous quotations from the literature. In contrast, use of a generic $\widehat{\boldsymbol{\beta}}$, as is often done, is ambiguous, on occasion as our $\widehat{\boldsymbol{\beta}}_{Rk}$, and other times as our $\widehat{\boldsymbol{\beta}}_c$. Failure to recognize these distinctions appears in part to explain entrapment in the critical errors of logic documented here, which we seek to avoid. Details emerge subsequently.

2.2 Canonical Form

Subsequent developments rest on the following. The singular decomposition $\boldsymbol{X} = \boldsymbol{P}\boldsymbol{D}_\xi \boldsymbol{Q}'$, with $\boldsymbol{P}'\boldsymbol{P} = \boldsymbol{I}_p$, together with $\boldsymbol{\theta} = \boldsymbol{Q}'\boldsymbol{\beta}$ as an orthogonal re-parametrization, give $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \rightarrow \boldsymbol{Y} = \boldsymbol{P}\boldsymbol{D}_\xi \boldsymbol{Q}'\boldsymbol{\beta} + \boldsymbol{\epsilon} \rightarrow \boldsymbol{U} = \boldsymbol{P}'\boldsymbol{Y} = \boldsymbol{D}_\xi\boldsymbol{\theta} + \boldsymbol{P}'\ \boldsymbol{\epsilon}$ as a succinct and equivalent model on $\mathbb{R}^p$. Gauss–Markov assumptions regarding $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ stipulate that $\mathrm{E}(\boldsymbol{\epsilon}) = \boldsymbol{0} \in \mathbb{R}^n$ and $\mathrm{V}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}_n$. From these it follows that $\mathrm{E}(\boldsymbol{P}'\boldsymbol{\epsilon}) = \boldsymbol{0} \in \mathbb{R}^p$ and $\mathrm{V}(\boldsymbol{P}'\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{P}'\boldsymbol{I}_n\boldsymbol{P} = \sigma^2\boldsymbol{I}_p$, so that $\mathrm{E}(\boldsymbol{U}) = \boldsymbol{D}_\xi\boldsymbol{\theta}$ and $\mathrm{V}(\boldsymbol{U}) = \sigma^2\boldsymbol{I}_p$ as in subsequent usage.

A critical property of residuals emerges subsequently. Beginning with $\widehat{\boldsymbol{\beta}}_R(k)$, define the residual sum of squares as $SS(k) = [\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_R(k)]'[\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_R(k)]$. We require the important result that

$$SS(k_1) < SS(k_2) \text{ for } 0 \leq k_1 < k_2. \tag{1}$$

On applying the foregoing canonical form, we find

$$\frac{dSS(k)}{dk} = 2\sum_{i=1}^{p} \frac{U_i^2\xi_i^2 k}{\left(\xi_i^2 + k\right)^3} > 0 \tag{2}$$

to complete the argument.

3 Ridge Regression

3.1 Essentials

We recall basics as set forth principally in Hoerl & Kennard (1970), Marquardt (1970), and Marquardt & Snee (1975). Following those authors, write the residual sum of squares as

$$\phi = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}) = \phi_{min} + \phi(\boldsymbol{B}), \tag{3}$$

where $\phi_{min} = (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_L)'(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_L)$ and $\phi(\boldsymbol{B}) = (\boldsymbol{B} - \widehat{\boldsymbol{\beta}}_L)'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{B} - \widehat{\boldsymbol{\beta}}_L)$.

Various assertions critical to the foundations of ridge regression have been made. These are enumerated here for subsequent reference and critique.

Assertions:

A1. Hoerl & Kennard (1970, p. 57), note that

$$\widehat{\boldsymbol{\beta}} = \left[\boldsymbol{I}_p + k(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]^{-1}\widehat{\boldsymbol{\beta}}_L \tag{4}$$

with $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{Rk}$.

A2. Hoerl & Kennard (1970, pp. 58–59): "The ridge trace can be shown to be following a path through the sums of squares surface so that for a fixed ϕ a single value for $\boldsymbol{B}$ is chosen and that is the one with minimal length." Precisely: Fix ϕ_0 and minimize $\boldsymbol{B}'\boldsymbol{B}$ subject to

$$(\boldsymbol{B} - \widehat{\boldsymbol{\beta}}_L)'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{B} - \widehat{\boldsymbol{\beta}}_L) = \phi_0. \tag{5}$$

Next proceed to minimize the LaGrangian function

$$F_0(\boldsymbol{B}, k) = \boldsymbol{B}'\boldsymbol{B} + (1/k)[(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}) - \phi_0]. \tag{6}$$

This reduces to

$$\boldsymbol{B} = \widehat{\boldsymbol{\beta}}_R(k_0) = \left(\boldsymbol{X}'\boldsymbol{X} + k_0\boldsymbol{I}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y} \tag{7}$$

where k_0 is chosen to satisfy the constraint (5).

A3. Hoerl & Kennard (1970, p. 59): "A completely equivalent statement of the problem is this: If the squared length of the regression vector $\boldsymbol{B}$ is fixed at R^2, then $\widehat{\boldsymbol{\beta}}$ is the value of $\boldsymbol{B}$ that gives a minimum sum of squares. That is, $\widehat{\boldsymbol{\beta}}$ is the value of $\boldsymbol{B}$ that minimizes the function

$$F_1(\boldsymbol{B}, k) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}) + k(\boldsymbol{B}'\boldsymbol{B} - R^2).\text{"} \tag{8}$$

Again the solution is $\boldsymbol{B} = \widehat{\boldsymbol{\beta}}_R(k_0) = (\boldsymbol{X}'\boldsymbol{X} + k_0\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{Y}$, choosing k_0 to satisfy the constraint $\{\boldsymbol{\beta}'\boldsymbol{\beta} = R^2\}$. Here we have replaced their $(1/k)$ by k to conform with conventional notation.

A4. (a) Hoerl & Kennard (1970, p. 59), regarding $F_0(\boldsymbol{B}, k)$: "Of course, in practice it is easier to choose a $k \geq 0$ and then to compute ϕ_0" at expression (5).

(b) Marquardt & Snee (1975, p. 5), regarding the dual $F_1(\boldsymbol{B}, k)$: "If $\widehat{\boldsymbol{\beta}}$ is the solution of $(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})\boldsymbol{\beta} = \boldsymbol{g}$, then $\widehat{\boldsymbol{\beta}}$ minimizes the sum of squares of residuals on the sphere centered at the origin whose radius is the length of $\widehat{\boldsymbol{\beta}}$." Here $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{Rk}$ and $\boldsymbol{g} = \boldsymbol{X}'\boldsymbol{Y}$.

3.2 A Critique

The foregoing assertions are critical, and their validity must be reassessed. Clearly $F_0(\boldsymbol{B}, k)$ and $F_1(\boldsymbol{B}, k)$ are in keeping with LaGrange's Principle; $F_1(\boldsymbol{B}, k)$ is dual to $F_0(\boldsymbol{B}, k)$; and solutions $\{\boldsymbol{\beta}, \lambda\}$ of $\{(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I}_p)\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}; \boldsymbol{\beta}'\boldsymbol{\beta} = c^2\}$ are necessary; moreover, that they are minimizing remains to be determined.

Assertion **A1:** True, but strictly for the unconstrained solutions $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{Rk}$. **A1** is false if misapplied taking $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_c$, as in section 3 of Hoerl & Kennard (1970). The fallacy is that since $\widehat{\boldsymbol{\beta}}_c$ necessarily lies on the sphere $\{\widehat{\boldsymbol{\beta}}_c'\widehat{\boldsymbol{\beta}}_c = c^2\}$, there is no one-to-one linear transformation taking $\widehat{\boldsymbol{\beta}}_L$ onto that sphere, so that $\widehat{\boldsymbol{\beta}}_c \neq [\boldsymbol{I}_p + k(\boldsymbol{X}'\boldsymbol{X})^{-1}]^{-1}\widehat{\boldsymbol{\beta}}_L$ as asserted. This error again appears to stem from ambiguous use of notation.

Assertion **A2:** Incomplete. It is necessary that (7) hold at a stationary point by LaGrange's Principle; however, the solution need not be minimizing. Specifically, for a solution of minimal length on the sum of squares surface at level ϕ_1, there may be a solution of the same length on the surface at level $\phi_0 < \phi_1$. If there are values $\phi_0 < \phi_1$ such that $\|\widehat{\boldsymbol{\beta}}_R(k_0)\| = \|\widehat{\boldsymbol{\beta}}_R(k_1)\|$ and $k_0 < k_1$, then $\widehat{\boldsymbol{\beta}}_R(k_0)$ must be preferred from (1) for its smaller residual $SS(k)$.

Assertion **A3:** Incomplete. From duality between (8) and (6), solutions constrained to the same length may occur at distinct values $k_0 < k_1$. Since $SS(k)$ is monotone increasing, the minimization of $F_1(\boldsymbol{B}, k)$ requires the choice of $\widehat{\boldsymbol{\beta}}_R(k_0)$ at the smaller value. Further details appear subsequently.

Assertions **A4:** False. Assertion (a) pertains to $F_0(\boldsymbol{B}, k)$ and (b) to its dual $F_1(\boldsymbol{B}, k)$. As in the Appendix, LaGrange minimization requires solving $(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I}_p)\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}$ (A.3) and $\{\boldsymbol{\beta}'\boldsymbol{\beta} = c^{*2}\}$ (A.4) for the $p + 1$ unknowns $(\beta_1, \ldots, \beta_p, k)$. Assertions **A4** are to fix $k = \ell > 0$; to solve (A.3) as $\widehat{\boldsymbol{\beta}}_{R\ell}$; and then to discover the implied constraint on evaluating $(\widehat{\boldsymbol{\beta}}_{R\ell} - \widehat{\boldsymbol{\beta}}_L)'\boldsymbol{X}'\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_{R\ell} - \widehat{\boldsymbol{\beta}}_L) = \phi_\ell$ at (5) under $F_0(\boldsymbol{B}, k)$, or evaluating $\{\widehat{\boldsymbol{\beta}}_{R\ell}'\widehat{\boldsymbol{\beta}}_{R\ell} = c_\ell^2\}$ under $F_1(\boldsymbol{B}, k)$, as appropriate. These assertions are tantamount to claiming that solutions of $\{(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I}_p)\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}\}$ tacitly embody the constraints (5) and $\{\boldsymbol{\beta}'\boldsymbol{\beta} = c^2\}$ as well, which clearly they cannot. Nonetheless, the solution $\widehat{\boldsymbol{\beta}}_{R\ell}$, of length c_ℓ, need not minimize the residual sum of squares $SS(k) = (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{Rk})'(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{Rk})$, as claimed. The fallacy stems from the false supposition that k and c^2 correspond one-to-one. To the contrary, in Section 3.3 we demonstrate solutions at different k's having the same length, say $\|\widehat{\boldsymbol{\beta}}_R(k_1)\| = \|\widehat{\boldsymbol{\beta}}_R(k_2)\|$ with $k_1 < k_2$. But then $\widehat{\boldsymbol{\beta}}_R(k_2)$ cannot be minimizing, as $SS(k_2) > SS(k_1)$ from monotonicity of $SS(k)$ as in (1). The dual assertions (a) and (b) often are misconstrued as equivalent statements regarding solutions $\widehat{\boldsymbol{\beta}}_{Rk}$ of (A.3) alone. See Nostrand (1980), for example. Remedial developments follow.

To continue, for fixed c define the set

$$\Lambda(c) = \{k : \| \widehat{\beta}_R(k) \| = c\}, \tag{9}$$

and let $k_c = \min\{\Lambda(c)\}$, as identified explicitly in Section 3.3. Then Assertion **A4**(b) may be amended as follows.

Assertion **A4*.** If $\widehat{\beta}_R(k)$ is a solution of $\{(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})\,\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}\}$ having length $\| \widehat{\beta}_R(k) \| = c^*$, then $\widehat{\beta}_R(k_c^*)$ minimizes the sum of squares of residuals on the sphere centered at the origin whose radius is the length c^* of $\widehat{\beta}_R(k)$, where $k_c^* = \min\{\Lambda(c^*)\}$.

Consequences of Assertions **A4** are profound: Their failure exposes an irreparable void in the foundations of ridge regression. Moreover, the implications of **A4*** are similarly far reaching. Of the many schemes on record for choosing k, it then is incumbent on the user to examine the set (9) for each such k. If a singleton set, then the solution is minimizing. That the algorithm **A4*** is required otherwise, has been neither recognized nor implemented heretofore. Further clarification emerges through a case study in Section 3.3.

A Referee draws parallel to Draper (1963) on "ridge analysis," as distinct from ridge regression, regarding (i) variational properties of second-order response surfaces on spheres of radius R in the predictor variables $\boldsymbol{x}$, and (ii) evolution of these as R varies. The basic relation is $(\boldsymbol{B} - \lambda\boldsymbol{I})\boldsymbol{x} = -\frac{1}{2}\boldsymbol{b}$, where λ is a LaGrange multiplier; $\boldsymbol{B}$ is the matrix of second-order coefficients, with pure quadratic terms on the diagonal; and $\boldsymbol{b}$ the vector of linear coefficients of the model. That λ and R are not one-to-one is noted by Draper (1963). The parallel, that k and c^2 are not one-to-one in ridge regression, becomes clear on comparing our (A.3) with the basic relation of Draper (1963) as cited.

3.3 Case Studies: Minimizing Solutions

We reexamine the Hospital Manpower Data as reported in Myers (1990). Records at $n = 17$ U. S. Naval Hospitals include: Y : Monthly man–hours; X_1 : Average daily patient load; X_2 : Monthly X–ray exposures; X_3 : Monthly occupied bed days; X_4 : Eligible population in the area divided by 1000; and X_5 : Average length of patients' stay in days. The basic model is

$$\{Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon_i;\ 1 \le i \le n\}. \tag{10}$$

Following Hoerl & Kennard (1970), Marquardt (1970), Marquardt & Snee (1975), Myers (1990), and others, we center and scale so that $\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\boldsymbol{Z}'\boldsymbol{Z}$ in "correlation form" having unit diagonals, the central focus being the rates of change $\boldsymbol{\beta} = [\beta_1, \beta_2, \beta_3, \beta_4, \beta_5]'$. The data are given in Table 3.8, pp. 132–133, of Myers (1990), and our computations rely heavily on PROC IML of the SAS Programming System. The data are exceedingly ill-conditioned, with singular values

$$\begin{aligned} \boldsymbol{D}_\xi &= \mathrm{Diag}(\xi_1, \xi_2, \xi_3, \xi_4, \xi_5) \\ &= \mathrm{Diag}(2.048687, 0.816997, 0.307625, 0.201771, 0.007347) \end{aligned}$$

for $\boldsymbol{Z}$.

Widely diverse criteria have evolved for choosing k, often with deeply diverging consequences regarding ridge estimators, predictors, and their properties. Five criteria in common use are reported in Table 1, together with definitions and their values for the Hospital Manpower Data. These include $DF_k = \mathrm{tr}(\boldsymbol{H}_k)$ with $\boldsymbol{H}_k = [\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z} + k\boldsymbol{I}_p)^{-1}\boldsymbol{Z}']$; the cross-validation $PRESS_k$ statistic of Allen (1974); a rotation-invariant *Generalized Cross Validation* (GCV_k) of Golub *et al.* (1979); C_k to achieve variance-bias trade-off as in Mallows (1973); and HKB_k as recommended by Hoerl *et al.* (1975) from simulations. As listed in Table 1, $SS_{Res,k}$ is the residual

Table 1
Choices for k in the Hospital Manpower Data corresponding to conventional criteria $DF_k, GCV_k, C_k, PRESS_k$, and HKB_k.

Name	Definition	Value for k
DF_k	$\mathrm{tr}(\boldsymbol{H}_k) = \sum_{i=1}^{p} \frac{\xi_i^2}{(\xi_i^2+k)}$	0.0004
GCV_k	$\frac{SS_{Res,k}}{[n-(1+\mathrm{tr}(\boldsymbol{H}_k))]^2}$	0.004787
C_k	$\frac{SS_{Res,k}}{\widehat{\sigma}^2} - n + 2 + 2\mathrm{tr}(\boldsymbol{H}_k)]$	0.0050
$PRESS_k$	$\sum_{i=1}^{n} e_{(i,k)}^2$	0.2300
HKB_k	$\frac{k\widehat{\sigma}^2}{\widehat{\boldsymbol{\beta}}'_L\widehat{\boldsymbol{\beta}}_L}$	0.616964

Table 2
Lengths of $\widehat{\boldsymbol{\beta}}_{Rk}$ in the Hospital Manpower Data, and mean root residuals $R(k) = [(\boldsymbol{Y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}_{Rk})'(\boldsymbol{Y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}_{Rk})]^{\frac{1}{2}}$, for designated values of k.

k	0.00	0.04	0.08	0.12	0.16	0.20	0.24	0.28
$\Vert\widehat{\boldsymbol{\beta}}_{Rk}\Vert$	394.67	137.82	33.14	31.50	70.02	99.19	122.10	140.73
$R(k)$	2,129.53	2,474.87	2,735.75	2,914.38	3,057.54	3,184.84	3,305.00	3,422.13
k	0.32	0.36	0.40	0.48	0.56	0.60	0.64	0.68
$\Vert\widehat{\boldsymbol{\beta}}_{Rk}\Vert$	156.25	169.40	180.69	199.00	213.09	218.93	224.11	228.70
$R(k)$	3,538.22	3,654.22	3,770.58	4,004.70	4,240.27	4,358.28	4,476.26	4,594.06
k	0.72	0.76	0.80	0.84	0.88	0.92	0.96	1.00
$\Vert\widehat{\boldsymbol{\beta}}_{Rk}\Vert$	232.79	236.42	239.65	242.53	245.08	247.34	249.33	251.09
$R(k)$	4,711.56	4,828.65	4,945.23	5,061.20	5,176.49	5,291.03	5,404.77	5,517.64

sum of squares using ridge regression; $\widehat{\sigma}^2$ is the *OLS* residual mean square; and $\{e_{(i,k)}^2\}$ are the *PRESS* residuals for ridge regression. Further details are given in Myers (1990, pp. 392–411), including numerical values for DF_k, C_k, and $PRESS_k$ as reported here. Further choices include $k \in \{0.01, 0.03, 0.05, 0.07, 0.09\}$ and others to be noted.

Are Solutions Minimizing? We have claimed Assertions **A4** to be false. Evidence is provided in Table 2, where lengths $\Vert\widehat{\boldsymbol{\beta}}_{Rk}\Vert$, and mean root residuals $R(k) = [(\boldsymbol{Y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}_{Rk})'(\boldsymbol{Y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}_{Rk})]^{\frac{1}{2}}$, are reported as k ranges over $[0, 1]$, as stipulated by Hoerl & Kennard (1970) and others when $\boldsymbol{Z}'\boldsymbol{Z}$ is in correlation form. Here $\widehat{\boldsymbol{\beta}}_{Rk} = [\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3, \widehat{\beta}_4, \widehat{\beta}_5]'$ are rates of change; similar trends are exhibited when $\boldsymbol{\beta}$ is expanded to include the intercept. It is seen that $\Vert\widehat{\boldsymbol{\beta}}_{Rk}\Vert$ decreases initially to a minimum, then increases beyond $k = 1.0$, but eventually decreases to zero through shrinkage.

In greater detail, the canonical form of Section 2.2 asserts that $\widehat{\boldsymbol{\beta}}_{Rk} = \boldsymbol{Q}\widehat{\boldsymbol{\theta}}_{Rk}$; that $\boldsymbol{Q}$ is orthogonal; and thus, letting $g_{\widehat{\beta}_R}(k) = \Vert\widehat{\boldsymbol{\beta}}_{Rk}\Vert^2$, that $g_{\widehat{\beta}_R}(k) = g_{\widehat{\theta}_R}(k)$. It follows that $g_{\widehat{\theta}_R}(k) = \sum_{i=1}^{p} U_i^2\xi_i^2/(\xi_i^2 + k)^2$. This is differentiable; its derivative is

$$\partial g_{\widehat{\theta}_R}(k)/\partial k = -2\sum_{i=1}^{p} U_i^2\xi_i^2\left(\xi_i^2 + k\right)^{-3}; \tag{11}$$

and its path traces evolution of the derivative as k varies. Moreover, as $k \downarrow 0$, we have $[\partial g_{\widehat{\theta}_R}(k)/\partial k]_{k=0} = -2\sum_{i=1}^{p} U_i^2/\xi_i^4$. Accordingly, in the Hospital Manpower Data the length of $\boldsymbol{\theta}_R(k)$ is seen to drop precipitously away from *OLS* at $k = 0$, since $\xi_p = \xi_5 = 0.007347$ and its fourth power appears in the denominator.

A detailed local view is provided in Table 3, to include not only $\Vert\widehat{\boldsymbol{\beta}}_{Rk}\Vert$ and $R(k)$, but also the ridge estimates $\widehat{\boldsymbol{\beta}}_{Rk} = [\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3, \widehat{\beta}_4, \widehat{\beta}_5]'$ in rows identified with k. Values of $\widehat{\boldsymbol{\beta}}_{Rk}$ for $k \in \{0.08, 0.11, 0.12\}$ duplicate those in Table 8.9 of Myers (1990), who reports ridge estimates for $k \in [0, 0.24]$ by increments of 0.01. It is seen that $\Vert\widehat{\boldsymbol{\beta}}_{Rk}\Vert$ takes its minimal value, 12.46150,

Table 3
Ridge estimators $\widehat{\beta}_{Rk}$, lengths of $\widehat{\beta}_{Rk}$, and mean root residuals $R(k) = [(Y - Z\widehat{\beta}_{Rk})'(Y - Z\widehat{\beta}_{Rk})]^{\frac{1}{2}}$, for designated values of k.

k	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_4$	$\widehat{\beta}_5$	$\Vert\widehat{\beta}_{Rk}\Vert$	$R(k)$
0.08	10.6354	0.065428	0.359139	6.3206	−30.7471	33.1448	2,735.75
0.08095	10.6118	0.065432	0.358279	6.3674	−28.9649	31.5000	2,740.68
0.08797	10.4475	0.065444	0.352298	6.6903	−16.4728	20.6250	2,775.83
0.0981	10.2378	0.065414	0.344681	7.0942	−0.3156	12.4645	2,823.03
0.09829	10.2342	0.065413	0.344548	7.1012	−0.0308	12.4615	2,823.89
0.0983	10.2340	0.065413	0.344541	7.1015	−0.0159	12.4615	2,823.93
0.11	10.0248	0.065325	0.336955	7.4935	16.3900	20.6251	2,874.22
0.12	9.8679	0.065217	0.331280	7.7785	28.8834	31.5000	2,914.38

at $k_{\min} = 0.09829$. Moreover, both $\widehat{\beta}_R(0.12)$ and $\widehat{\beta}_R(0.08095)$ have the same length, namely 31.500, so that $\Lambda(31.500) = \{0.08095, 0.12\}$ as in (9). Were a user to choose $\widehat{\beta}_R(0.12)$ from Table 3, then this solution is not minimizing of length 31.500, as may be seen from $R(0.12) = 2{,}914.38 > 2{,}740.68 = R(0.08095)$. It is likewise clear that $\Lambda(20.625) = \{0.08797, 0.11\}$, and that $\widehat{\beta}_R(0.11)$ of Table 3 is not minimizing, to be replaced instead by $\widehat{\beta}_R(0.08797)$. Further violations of Assertion **A4**(b) follow on reflecting k asymmetrically about $k_{\min} = 0.09829$: The smaller k of each pair designates the minimizing solution, to demonstrate how Assertion **A4*** may be implemented.

To continue regarding Assertion **A4**(a), we have found $\phi_0 = 2{,}740.68^2 < 2{,}914.38^2 = \phi_1$ such that the corresponding minimal solutions of $F_0(\boldsymbol{B}, k)$ are $\widehat{\beta}_R(0.08095)$ and $\widehat{\beta}_R(0.12)$, having identical squared lengths. The preferred solution necessarily is $\widehat{\beta}_R(0.08797)$ with smaller residual sum of squares $SS(k)$. To summarize, in regard to **A4**(b) we have just seen that there are two solutions of $F_1(\boldsymbol{B}, k)$ at $\Vert \widehat{\beta}_{Rk} \Vert = 31.500$, namely, those corresponding to $k \in \Lambda(31.500) = \{0.08095, 0.12\}$ from (9). Clearly some values for k are inadmissible.

Specifically, admissible values for k, for which solutions of given length c^* are minimizing, are subject to rigid but yet undocumented constraints. To fix ideas, suppose that $\{g_{\widehat{\beta}_R}(0.00) > c^{*2} \geq c^2 \geq g_{\widehat{\beta}_R}(0.09829) = 12.46150^2 = 155.2877\}$. Then the only feasible values for k are in $[\min g_{\widehat{\beta}_R}^{-1}(c^2), 0.09829]$. For example, for c^2 in $\{33.14481^2 = 1{,}098.5784 \geq c^2 \geq 155.2877\}$, the feasible values from Table 3 are $k \in [0.08, 0.09829]$. Instead, for c^2 in $\{g_{\widehat{\beta}_R}(0.00) \geq c^2 \geq g_{\widehat{\beta}_R}(0.09829)\}$, the feasible values are $k \in [0.00, 0.09829]$. These are the only feasible values for $k \in [0, 1]$, so that $PRESS_k$ and HKB_k, as reported in Table 1, are not viable choices whenever $c^2 \geq 155.2877$. Conversely, choosing $\{0 < c^2 < g_{\widehat{\beta}_R}(0.09829) = 155.2877\}$ requires k in the interval $(\min g_{\widehat{\beta}_R}^{-1}(c^2), \infty)$, where $\min g_{\widehat{\beta}_R}^{-1}(155.2877) > 158$. Thus for $c^2 < 100$, the feasible values are $k \in (198, \infty)$, far surpassing the recommended interval $[0, 1]$, so that constraints $c^{*2} \in (0, 155.2877)$ are inadmissible in the Hospital Manpower Data. Values reported for $\{0 < c^2 < g_{\widehat{\beta}_R}(0.09829) = 155.2877\}$ and $\min g_{\widehat{\beta}_R}^{-1}(155.2877) > 158$ are supported by Maple software.

Origins of these anomalies may be traced as follows: (i) Evolution of $|\widehat{\beta}_5(k)|$ is down-up-down, beginning with $\widehat{\beta}_5(0.00) = -394.3280$, decreasing to zero between $k = 0.09$ and $k = 0.10$, increasing to $\widehat{\beta}_5(1.00) = 250.8307$ and beyond, but ultimately vanishing through shrinkage. (ii) $|\widehat{\beta}_5(k)|$ dominates other estimates by factors ranging from one to four except near its minimum. (iii) Other coefficients exhibit much narrower ranges for $k \in [0, 1]$. (iv) In consequence, $\Vert \widehat{\beta}_{Rk} \Vert$ is determined largely by $|\widehat{\beta}_5(k)|$. That $\Lambda(c^*)$ takes two values in these examples may be attributed to the down-up-down character of $\Vert\widehat{\beta}_{Rk}\Vert$. More elements could emerge in $\Lambda(c^*)$, for example, were a dominant estimate to undergo multiple sign changes, and other coefficients change signs as well. Such properties of ridge traces are studied in Zhang & McDonald (2005), and references cited, under further structure of $\boldsymbol{Z}'\boldsymbol{Z}$. Sign changes, crossings, rates-of-change of individual ridge estimates, and bounds on the number of sign changes, are determined by those authors on identifying zeros and derivatives of polynomials in k of degree $k - 1$. Those studies in turn suggest the cardinality of $\Lambda(c^*)$ at (9) to be finite.

In short, imbedded in the Hospital Manpower Data are hidden feasible constraints $\{\boldsymbol{\beta}'\boldsymbol{\beta} = c^2\}$ with $c^2 \geq 155.2877$. These would remain hidden, as in the past four decades, absent the foregoing detailed analyses. Corresponding constraints in other data sets may be unmasked accordingly.

3.4 Ridge in Retrospect

Many claims in ridge regression are attributed mistakenly to constrained solutions. Examples follow in varying detail.

The Variance-Bias Tradeoff. Hoerl & Kennard (1970, Section 4), purport (i) to assess variance and bias of $\widehat{\boldsymbol{\beta}}$, asserted in Section 3 as subject to $\{\boldsymbol{\beta}'\boldsymbol{\beta} = c^2\}$, and (ii) to prove in theorem 4.3 existence of $k > 0$ such that $\widehat{\boldsymbol{\beta}}$ dominates $\widehat{\boldsymbol{\beta}}_L$ in expected mean square (*EMS*). They claim the *EMS* decomposition $\mathrm{E}[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})] = \gamma_1(k) + \gamma_2(k)$, identifying $\gamma_1(k)$ as the sum of variances, and $\gamma_2(k)$ as "the square of a bias introduced when $\widehat{\boldsymbol{\beta}}$ is used rather than $\widehat{\boldsymbol{\beta}}_L$." Graphs of $\gamma_1(k)$, decreasing in k, of $\gamma_2(k)$, increasing in k, and of $\gamma_1(k) + \gamma_2(k)$, are reported for $k \in [0, 1]$ from a case study by those authors. The foregoing claims, including theorem 4.3, are valid only for the unconstrained ridge solutions, i.e. for $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{Rk}$ satisfying (A.3) alone, but not for $\widehat{\boldsymbol{\beta}}_c$ as intended. The facts diverge critically when constrained solutions $\widehat{\boldsymbol{\beta}}_c$ are taken properly into account.

Specifically, for $\{\widehat{\boldsymbol{\beta}}_c'\widehat{\boldsymbol{\beta}}_c = c^2\}$ let $\mathrm{E}(\widehat{\boldsymbol{\beta}}_c) = \boldsymbol{\beta}_c = [\beta_{c1}, \beta_{c2}, \ldots, \beta_{cp}]'$; then $\boldsymbol{\beta}_c$ is in the interior of $B_c^o = \{\boldsymbol{u} \in \mathbb{R}^p : \boldsymbol{u}'\boldsymbol{u} \leq c^2\}$ from convexity of $\mathrm{E}\,(\cdot)$. Since $\mathrm{E}(\widehat{\beta}_{ci}^2) = \mathrm{var}(\widehat{\beta}_{ci}) + \beta_{ci}^2$ and $\mathrm{E}(\widehat{\boldsymbol{\beta}}_c'\widehat{\boldsymbol{\beta}}_c) = \sum_{i=1}^p \mathrm{var}(\widehat{\beta}_{ci}) + \boldsymbol{\beta}_c'\boldsymbol{\beta}_c = c^2$ is fixed, it follows that $\sum_{i=1}^p \mathrm{var}(\widehat{\beta}_{ci}) = c^2 - \boldsymbol{\beta}_c'\boldsymbol{\beta}_c > 0$. Accordingly, on small spheres the total variance is necessarily small, but increasing with c, contradicting for $\widehat{\boldsymbol{\beta}}_c$ that $\gamma_1(k)$ is decreasing. For example, $\sum_{i=1}^p \mathrm{var}(\widehat{\beta}_{ci}) = c^2$ when $\mathcal{L}(\widehat{\boldsymbol{\beta}}_c)$ is uniform on the c-sphere. Similarly, the squared bias $E[(\widehat{\boldsymbol{\beta}}_c - \boldsymbol{\beta})'(\widehat{\boldsymbol{\beta}}_c - \boldsymbol{\beta})]$ need not be monotone, as asserted for $\gamma_2(k)$. To continue, a collection of *ridge traces,* as graphs of the p individual components of $\widehat{\boldsymbol{\beta}}_{Rk}$ as k evolves, are advocated by Hoerl & Kennard (1970) for use in choosing k. Following Hoerl & Kennard (1970), the *EMS* decomposition $\mathrm{E}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \gamma_1(k) + \gamma_2(k)$ is always depicted in the literature as continuous in k, as are the ridge traces. However, neither can be continuous in the case of $\widehat{\boldsymbol{\beta}}_c$, owing to admissibility constraints on k as in Section 3.3. These lapses appear to be further detritus stemming from ambiguity in the use of $\widehat{\boldsymbol{\beta}}$, since asserted properties were derived under the mistaken notion that constrained solutions are linear in $\widehat{\boldsymbol{\beta}}_L$.

Other Anomalies. Numerous authors, *expressly* concerned with constrained minimization, take solutions $\widehat{\boldsymbol{\beta}}_R = \{\widehat{\boldsymbol{\beta}}_{Rk} = (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I}_p)^{-1}\boldsymbol{X}'\boldsymbol{Y}; k \geq 0\}$, solving (A.3) alone, as *the* estimators in ridge regression, despite $\widehat{\boldsymbol{\beta}}_c$ or $\widehat{\boldsymbol{\beta}}_c^o$ as constrained minimizers. Examples are Hoerl & Kennard (1970), Marquardt (1970), Marquardt & Snee (1975), Golub *et al.* (1979), and others. In short, ridge regression rests on $\{\widehat{\boldsymbol{\beta}}_{Rk}; k \geq 0\}$ essentially through an accident of history.

Specifically, expressions for variances and biases; solutions for k purporting to minimize expected mean squares; prediction, validation, and cross-validation; and other aspects of ridge regression: All are predicated on the unconstrained solutions $\{\widehat{\boldsymbol{\beta}}_{Rk}; k \geq 0\}$ and their linearity in $\widehat{\boldsymbol{\beta}}_L$. If instead either $\{\widehat{\boldsymbol{\beta}}_c;\ c > 0\}$ or $\{\widehat{\boldsymbol{\beta}}_c^o;\ c > 0\}$ were taken, in keeping with proper LaGrange constraints, then the ensuing "ridge regressions" would differ dramatically, both *inter se,* and from extant results based on $\{\widehat{\boldsymbol{\beta}}_{Rk}; k \geq 0\}$. These discrepancies necessarily would encompass (i) the stability of the solutions $\widehat{\boldsymbol{\beta}}_c$ or $\widehat{\boldsymbol{\beta}}_c^o$ instead of $\widehat{\boldsymbol{\beta}}_{Rk}$, in comparison with $\widehat{\boldsymbol{\beta}}_L$; (ii) the inflation or deflation of variances, taking into account actual moments to be reckoned from constrained distributions; (iii) prediction using $\widehat{\boldsymbol{Y}}_c = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_c$ or $\widehat{\boldsymbol{Y}_c^o} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_c^o$, instead of $\widehat{\boldsymbol{Y}_{Rk}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{Rk}$; (iv) the use, meaning, and properties of cross-validative and predictive criteria; (v) discarding or rethinking $\{DF_k, GCV_k, C_k, PRESS_k, HKB_k\}$ and other means for choosing a particular solution; (vi) ridge traces, as modified to take into account $\{\widehat{\beta}^o_c; c > 0\}$ and singularity of the joint distribution of $\{\widehat{\beta}_c; c > 0\}$ for fixed c; and (vii) the trade-off between bias and variance of the constrained estimators $\widehat{\beta}_c$ and $\widehat{\beta}^o_c$, to be determined using their actual moments. Other discrepancies may be noted. All such issues beg to be examined anew, if indeed constrained optimization is to remain a viable foundation. However, any such re-examination might be preempted by insurmountable intricacies owing to discontinuities in admissible values for k, and to the corresponding nonstandard distributions. These examples demonstrate unequivocally that little of ridge regression, as currently practiced, rests on constrained solutions.

Extensive simulations, beyond the scope of this review but exemplified in Dempster *et al.* (1977), have compared numerous ridge and shrinkage solutions *inter se* and with *OLS*. The meaning, and the practical implications of all such studies, are largely obscured by the fact that the solutions $\widehat{\boldsymbol{\beta}}_{Rk}$ need not be minimizing. Careful reworking would require algorithms based on Assertion **A4*** together with admissible values for k.

By analogy to earlier sections, Hoerl & Kennard (1970, Section 8) posed generalized ridge solutions invoking the p equations $(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{K})\,\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}$, with non-negative ridge parameters $\boldsymbol{K} = \text{Diag}(k_1, \ldots, k_p)$. This, too, is incompatible with $\{\beta_1^2 = c_1^2, \ldots, \beta_p^2 = c_p^2\}$, in continuation of their equality constraints, for then the data can serve only to determine signs of the roots $\{\widehat{\beta}_1 = \pm c_1, \ldots, \widehat{\beta}_p = \pm c_p\}$. On the other hand, correct solutions constrained by $\{\beta_1^2 \le c_1^2, \ldots, \beta_p^2 \le c_p^2\}$ are provided by Myoken & Uchida (1977), akin to those of Balakrishnan (1963) where $\{k_1 = \cdots = k_p = k\}$.

4 Foundations via Conditioning

We have seen that ridge solutions $\{\widehat{\boldsymbol{\beta}}_{Rk}; k > 0\}$ are neither constrained nor necessarily minimizing, and therefore little more than *ad hoc* contrivances. We turn instead to conditioning of the linear systems $\{\widehat{\boldsymbol{\beta}}_{Rk}; k \ge 0\}$ for a re-appraisal of ridge regression and a promising alternative. As in Section 3.3, the model $\{\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}\}$ often is featured after scaling and centering $\boldsymbol{X}$ as a first step towards improved conditioning. Much that follows pertains to both.

4.1 Ills of Ill-Conditioning

Ill-conditioning of $\boldsymbol{Y} = \boldsymbol{X}\beta + \epsilon$ is characterized by (i) non-orthogonality of columns of $\boldsymbol{X}$; (ii) excessive lengths of $\widehat{\boldsymbol{\beta}}_L$; (iii) excessive dispersion parameters $\text{V}(\widehat{\boldsymbol{\beta}}_L)$; and (iv) instability of solutions $\widehat{\boldsymbol{\beta}}_L(\boldsymbol{X}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ owing to disturbances in $\boldsymbol{X}$, when considered as a transformation of $\boldsymbol{Y}$. Let $\boldsymbol{W} = \boldsymbol{X}'\boldsymbol{X}$ and $\boldsymbol{V} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$. A number of concepts have emerged.

(i) **Conditioning of $\boldsymbol{X}$**: Non-orthogonality of columns of $\boldsymbol{X}$ is gauged directly through condition numbers $c_\phi(\boldsymbol{X})$, and indirectly through $c_\phi(\boldsymbol{X}'\boldsymbol{X})$, to be defined.

(ii) **Excessive Length:** Hoerl & Kennard (1970, p. 57), correctly note for $k > 0$ the attenuation $\widehat{\boldsymbol{\beta}}'_{Rk}\widehat{\boldsymbol{\beta}}_{Rk} < \widehat{\boldsymbol{\beta}}'_L\widehat{\boldsymbol{\beta}}_L$ of ridge solutions relative to *OLS*. These accordingly are called "shrinkage estimators."

(iii) **Variance Inflation:** Since $\text{V}(\widehat{\boldsymbol{\beta}}_L) = \sigma^2\boldsymbol{V}$, the *variance inflation factors* (*VIF*s) of $\widehat{\boldsymbol{\beta}}_L = [\widehat{\beta}_{L_1}, \ldots, \widehat{\beta}_{L_p}]'$ are defined as $\{VIF(\widehat{\beta}_{L_j}) = v_{jj}/w_{jj}^{-1}; 1 \le j \le p\}$, i.e. the ratio of the actual to the "ideal" variance attained when columns of $\boldsymbol{X}$ are orthogonal, in which case $\boldsymbol{W} = \text{Diag}\,(w_{11}, \ldots, w_{pp})$. If instead $\{\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\beta} + \epsilon\}$, with $\boldsymbol{Z}'\boldsymbol{Z}$ in "correlation form," then $\boldsymbol{V} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}$ and $\{VIF(\widehat{\beta}_{L_j}) = v_{jj}; 1 \le j \le p\}$ from scale-invariance of *VIF*s. Designate

$\{V_1 \geq V_2 \geq \cdots \geq V_p\}$ as the *ordered* diagonal elements of $\boldsymbol{V} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}$; Marquardt & Snee (1975) identify V_1 to be "the best single measure of the conditioning of the data," thus a critical diagnostic tool. See also Marquardt (1970), Beaton *et al.* (1976), and Davies & Hutton (1975), for example. To continue, following standard usage in ridge regression, define the condition number $c_1(\boldsymbol{Z}'\boldsymbol{Z}) = \xi_1^2/\xi_p^2$, with $\{\xi_1^2 \geq \xi_2^2 \geq \cdots \geq \xi_p^2 > 0\}$ as the ordered eigenvalues of $\boldsymbol{Z}'\boldsymbol{Z}$. A basic connection between *VIF*s and condition numbers is due to Berk (1977): $V_1 \leq c_1(\boldsymbol{Z}'\boldsymbol{Z}) \leq p(V_1 + \cdots + V_p)$. Since $c_1(\boldsymbol{A}) = c_1(\boldsymbol{A}^{-1})$, Berk's result may be expressed equivalently for *OLS* as

$$V_1 \leq c_1[\mathrm{V}(\widehat{\boldsymbol{\beta}}_L)] \leq p\,\mathrm{tr}[\mathrm{V}(\widehat{\boldsymbol{\beta}}_L)], \tag{12}$$

a statement about conditioning of $\mathrm{V}(\widehat{\boldsymbol{\beta}}_L)$.

(iv) Stability and Conditioning: Stability of the system $\boldsymbol{Az} = \boldsymbol{b}$ refers to the propagation of small perturbations in elements of $\boldsymbol{A}$ to disturbances in the solution. This is gauged by the *condition number* $c_g(\boldsymbol{A}) = g(\boldsymbol{A})g(\boldsymbol{A}^{-1})$ when $\boldsymbol{A}$ is invertible, where $g(\cdot)$ ordinarily is a norm. A system is well conditioned at $\boldsymbol{A} = \boldsymbol{I}_p$ with $c_g(\boldsymbol{I}_p) = 1.0$, larger values reflecting greater ill-conditioning. In particular, $c_1(\boldsymbol{A})$ as defined belongs to the class $\{c_\phi(\cdot);\ \phi \in \Phi\}$ generated by the matrix norms invariant under orthogonal congruence; see Marshall & Olkin (1979) and Horn & Johnson (1985). Since for $k \geq 0, c_\phi(\boldsymbol{X}'\boldsymbol{X}) \geq c_\phi(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I}_p)$ improves conditioning, Marshall & Olkin (1979, p. 273), cite this as justification for taking $\{\boldsymbol{X}'\boldsymbol{X}\beta = \boldsymbol{X}'\boldsymbol{Y}\} \rightarrow \{(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I}_p)\,\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}\}$ in ridge regression.

If instead $\boldsymbol{Az} = \boldsymbol{b}$ is consistent, $\boldsymbol{A}$ is non-invertible and not necessarily square, we have $c_g(\boldsymbol{A}) = g(\boldsymbol{A})g(\boldsymbol{A}^{\dagger})$ with $\boldsymbol{A}^{\dagger}$ as the Moore–Penrose inverse, whereas $c_1(\boldsymbol{A})$ now is the ratio of its largest to smallest singular value. To study the stability of solutions to disturbances in $\boldsymbol{X}$, begin with $\widehat{\boldsymbol{\beta}}_L(\boldsymbol{X}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$; observe that conditioning of $[(\boldsymbol{X}'\boldsymbol{X})^{-1}\,\boldsymbol{X}']$ gauges the propagation of disturbances in $\boldsymbol{X}$ to the solution $\widehat{\boldsymbol{\beta}}_L$; infer from its canonical form that $\widehat{\boldsymbol{\beta}}_L(\boldsymbol{X}) = \boldsymbol{Q}\widehat{\boldsymbol{\theta}}_L(\boldsymbol{D}_\xi)$ with $\widehat{\boldsymbol{\theta}}_L(\boldsymbol{D}_\xi) = \boldsymbol{D}_\xi^{-2}\boldsymbol{D}_\xi\boldsymbol{U} = \boldsymbol{D}_\xi^{-1}\boldsymbol{U}$ as in Section 2.2; and, from the orthogonal invariance of condition numbers, infer that stability of $\widehat{\boldsymbol{\beta}}_L(\boldsymbol{X})$ is gauged through conditioning by

$$c_\phi[\widehat{\boldsymbol{\beta}}_L(\boldsymbol{X})] = c_\phi[\widehat{\boldsymbol{\theta}}_L(\boldsymbol{D}_\xi)] = c_\phi\left(\boldsymbol{D}_\xi^{-1}\right) = c_\phi(\boldsymbol{X}). \tag{13}$$

This approach conforms with that of Belsley *et al.* (1980). After scaling columns of $\boldsymbol{X}$ to $\boldsymbol{Z}$ having (approximately) equal lengths and improved conditioning, those authors invoked the concept of *elasticities* in order to link sensitivities of solutions $\widehat{\boldsymbol{\beta}}_L = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}$, and their variances, to disturbances in $\boldsymbol{Z}$, as gauged by its condition number $c_1(\boldsymbol{Z})$.

4.2 Critical Issues

Ridge regression seeks redress for deficiencies in *OLS*. By the same standards, it is essential to gauge progress of ridge regression towards these goals, a matter largely unaddressed in the literature. Clearly the system $\{(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I}_p)\,\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}\}$ itself is not orthogonal: The solutions $\widehat{\boldsymbol{\beta}}_{Rk}$ are themselves subject to variance inflation; their *VIF*s necessarily exceed unity; and these quantities continue to remain at issue for $\widehat{\boldsymbol{\beta}}_{Rk}$. The capacity of ridge regression to ameliorate the excessive dispersion of *OLS* thus hinges on improving *VIF*s and condition numbers for $\mathrm{V}(\widehat{\boldsymbol{\beta}}_{Rk})$. Similarly, the stability of ridge solutions rests on conditioning of $\widehat{\boldsymbol{\beta}}_{Rk}(\boldsymbol{X}) = (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I}_p)^{-1}\boldsymbol{X}'\boldsymbol{Y}$ to perturbations in $\boldsymbol{X}$ when considered as a data transformation, i.e. on $c_1\ [(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I}_p)^{-1}\ \boldsymbol{X}']$. These concerns for $\widehat{\boldsymbol{\beta}}_{Rk}$ emerge precisely for the same reasons as for *OLS*, prompting questions to be considered subsequently. Silence on these matters may attribute to the following dismissive assessment of the system $\{(\boldsymbol{Z}'\boldsymbol{Z} + k\boldsymbol{I}_p)\,\boldsymbol{\beta} = \boldsymbol{Z}'\boldsymbol{Y}\}$ by Hoerl & Kennard (1970, p. 65): "At a certain value of k the system will stabilize and have the general characteristics of an orthogonal system." This assertion remains to be examined, seeming more plausible with increasing k. Questions pertaining to these matters are listed next.

- **Q1**: Does the system stabilize sufficiently for $k \in (0,1]$ that *VIF*s for elements of $\widehat{\boldsymbol{\beta}}_{Rk}$ are no longer excessive?
- **Q2**: With regard to $\mathrm{V}(\widehat{\boldsymbol{\beta}}_{Rk}) = \sigma^2(\boldsymbol{Z}'\boldsymbol{Z} + k\boldsymbol{I}_p)^{-1}\boldsymbol{Z}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z} + k\boldsymbol{I}_p)^{-1}$, is its condition number $c_1[\mathrm{V}(\widehat{\boldsymbol{\beta}}_{Rk})]$ diminished and stable at some $k \in (0,1]$ in comparison with *OLS*?
- **Q3** Do the solutions $\widehat{\boldsymbol{\beta}}_{Rk}(\boldsymbol{X})$ stabilize as data transformations owing to disturbances in $\boldsymbol{X}$ for some $k \in (0,1]$?
- **Q4** Does $\mathrm{E}(\widehat{\boldsymbol{\beta}}_{Rk}) = T(\boldsymbol{\beta})$ become well-conditioned as a parameter transformation for some $k \in (0,1]$?

Regarding the latter, the conditioning of $\mathrm{E}(\widehat{\boldsymbol{\beta}}_{Rk}) = T(\boldsymbol{\beta})$ is at issue in transforming parameters, as in assessing the trade-off between variance and bias. Observe for completeness that the listed items pertain not only to ridge estimators $\{\widehat{\boldsymbol{\beta}}_{Rk}; k \ge 0\}$, but also to other biased solutions including $\{\widehat{\boldsymbol{\beta}}_{Sk}; k \ge 0\}$ of Section 4.4.

4.3 Ridge Regression

Taking $\boldsymbol{W} = \boldsymbol{X}'\boldsymbol{X} \to (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I}_p)$ informally serves to ease apparent dependencies among columns of $\boldsymbol{X}$ when ill-conditioned, in that elements $\{r_{ij} = w_{ij}/\sqrt{w_{ii}w_{jj}}\} \to \{r_{ij}(k) = w_{ij}/\sqrt{(w_{ii}+k)(w_{jj}+k)}\}$ decrease in magnitude with increasing k. More formally, that $\{\boldsymbol{X}'\boldsymbol{X} \to (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I}_p)\}$ explicitly reduces condition numbers for $k > 0$ is cited by Marshall & Olkin (1979) as grounds for ridge regression, as noted. Moreover, for condition numbers $c_1(\cdot)$, the improvement is quantified directly on comparing $c_1(\boldsymbol{X}'\boldsymbol{X}) = \xi_1^2/\xi_p^2$ with $c_1(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I}_p) = (\xi_1^2 + k)/(\xi_p^2 + k)$, where $\{\xi_1 \ge \xi_2 \ge \cdots \ge \xi_p\}$ are singular values of $\boldsymbol{X}$.

4.4 Surrogate Models

Closer inspection sees the correspondence $\boldsymbol{A}\boldsymbol{z} = \boldsymbol{b} \longleftrightarrow \boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}$ fail in linear inference, since $\boldsymbol{A} = \boldsymbol{X}'\boldsymbol{X}$ and $\boldsymbol{b} = \boldsymbol{X}'\boldsymbol{Y}$ both are subject to disturbances in $\boldsymbol{X}$, rather than disturbances only in $\boldsymbol{A}$. While $\{\boldsymbol{X}'\boldsymbol{X} \to (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I}_p)\}$ does improve conditioning on the left, the conditioning of $\boldsymbol{X}$, and its consequent effects on solutions, persists on the right in $\{(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I}_p)\ \boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}\}$. To continue, observe that $\{\boldsymbol{X}'\boldsymbol{X} \to (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I}_p)\}$ is tantamount to modifying $\boldsymbol{X}$ itself as a means to enhanced conditioning. Specifically, the singular decomposition $\boldsymbol{X} = \boldsymbol{P}\boldsymbol{D}_\xi\ \boldsymbol{Q}'$, and the recomposition $\boldsymbol{X}_k = \boldsymbol{P}\mathrm{Diag}((\xi_1^2 + k)^{\frac{1}{2}}, \ldots, (\xi_p^2 + k)^{\frac{1}{2}})\boldsymbol{Q}'$, give $\boldsymbol{X}_k^{'}\boldsymbol{X}_k = (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I}_p)$; these together show that ridge regression entails the system $\{\boldsymbol{X}_k^{'}\boldsymbol{X}_k\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}\}$. Instead, we take $\{\boldsymbol{Y} = \boldsymbol{X}_k\boldsymbol{\beta} + \boldsymbol{\epsilon}\}$ as an approximation, or *surrogate,* for the ill-conditioned model $\{\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}\}$ itself, as in the following.

Definition 1: Given an ill-conditioned model $\{\boldsymbol{Y} = \boldsymbol{X}\beta + \epsilon\}$, its *ridge surrogate* is a modified model $\{\boldsymbol{Y} = \boldsymbol{X}_k\boldsymbol{\beta} + \boldsymbol{\epsilon}\}$. The *surrogate estimator* $\widehat{\boldsymbol{\beta}}_{Sk}$, solving $\boldsymbol{X}_k^{'}\boldsymbol{X}_k\boldsymbol{\beta} = \boldsymbol{X}_k^{'}\boldsymbol{Y}$, is *OLS* for the surrogate model. Moreover, the order of approximation of $\boldsymbol{X}_k$ for $\boldsymbol{X}$ may be gauged by the squared Frobenius distance $\|\boldsymbol{X} - \boldsymbol{X}_k\|_F^2 = \sum_{i=1}^{p}(\xi_i - \sqrt{\xi_i^2 + k})^2$.

We motivate these developments through condition numbers. Take $\boldsymbol{X}$ into its column-scaled form $\boldsymbol{Z}$; let $\{\xi_1 \ge \xi_2 \ge \cdots \ge \xi_p > 0\}$ be ordered singular values for $\boldsymbol{Z}$; let $\boldsymbol{Z}_k = \boldsymbol{P}\mathrm{Diag}((\xi_1^2 + k)^{\frac{1}{2}}, \ldots, (\xi_p^2 + k)^{\frac{1}{2}})\boldsymbol{Q}'$, and invoke elasticities as in Belsley *et al.* (1980).

Table 4
Properties of canonical estimators $\{\widehat{\boldsymbol{\theta}}_L, \widehat{\boldsymbol{\theta}}_{Rk}, \widehat{\boldsymbol{\theta}}_{Sk}\}$, *where* $\widehat{\boldsymbol{\theta}} = \boldsymbol{DU} = \boldsymbol{DP'Y}$ *and* $\boldsymbol{D}(\omega_i) = \mathrm{Diag}(\omega_1, \ldots, \omega_p)$.

Estimator	$\boldsymbol{D}$	$\mathrm{E}(\widehat{\boldsymbol{\theta}}) = T(\boldsymbol{\theta})$	$\mathrm{V}(\widehat{\boldsymbol{\theta}})$
$\widehat{\boldsymbol{\theta}}_L = \boldsymbol{D}_L\boldsymbol{U}$	$\boldsymbol{D}_\xi^{-1}$	$\boldsymbol{\theta}$	$\sigma^2\boldsymbol{D}_\xi^{-2}$
$\widehat{\boldsymbol{\theta}}_{Rk} = \boldsymbol{D}_R\boldsymbol{U}$	$\boldsymbol{D}\left(\frac{\xi_i}{\xi_i^2+k}\right)$	$\boldsymbol{D}\left(\frac{\xi_i^2}{\xi_i^2+k}\right)\boldsymbol{\theta}$	$\sigma^2\boldsymbol{D}\left(\frac{\xi_i^2}{(\xi_i^2+k)^2}\right)$
$\widehat{\boldsymbol{\theta}}_{Sk} = \boldsymbol{D}_S\boldsymbol{U}$	$\boldsymbol{D}\left(\frac{1}{\sqrt{\xi_i^2+k}}\right)$	$\boldsymbol{D}\left(\frac{\xi_i}{\sqrt{\xi_i^2+k}}\right)\boldsymbol{\theta}$	$\sigma^2\boldsymbol{D}\left(\frac{1}{\xi_i^2+k}\right)$

We now may link the sensitivities of solutions $\widehat{\boldsymbol{\beta}}_{Sk} = (\boldsymbol{Z}_k'\boldsymbol{Z}_k)^{-1}\boldsymbol{Z}_k\boldsymbol{Y}$, and of their variances, to the condition number $c_1(\boldsymbol{Z}_k) = (\xi_1^2 + k)^{\frac{1}{2}}/(\xi_p^2 + k)^{\frac{1}{2}}$. It remains to study properties of the surrogate estimators $\{\widehat{\boldsymbol{\beta}}_{Sk}; k \geq 0\}$, in comparison with *OLS* and ridge solutions $\{\widehat{\boldsymbol{\beta}}_{Rk}; k \geq 0\}$.

4.5 Properties of $\widehat{\boldsymbol{\beta}}_{Rk}$ and $\widehat{\boldsymbol{\beta}}_{Sk}$

Direct comparisons among $\{\widehat{\boldsymbol{\beta}}_L, \widehat{\boldsymbol{\beta}}_{Rk}, \widehat{\boldsymbol{\beta}}_{Sk}\}$ become more transparent in the canonical forms of Section 2.2. Specifically, arguments leading to (13) demonstrate that $\widehat{\boldsymbol{\theta}}_L = \boldsymbol{D}_\xi^{-1}\boldsymbol{U}$, $\mathrm{E}(\widehat{\boldsymbol{\theta}}_L) = \boldsymbol{\theta}$, and $\mathrm{V}(\widehat{\boldsymbol{\theta}}_L) = \sigma^2\boldsymbol{D}_\xi^{-2}$ under *OLS*, as reported in Table 4 under Gauss–Markov assumptions. Parallel developments for the ridge estimators $\{\widehat{\boldsymbol{\theta}}_{Rk}; k \geq 0\}$, and the surrogate ridge estimators $\{\widehat{\boldsymbol{\theta}}_{Sk}; k \geq 0\}$, each in canonical form, are reported also in Table 4. Both $\{\widehat{\boldsymbol{\theta}}_{Rk}; k \geq 0\}$ and $\{\widehat{\boldsymbol{\theta}}_{Sk}; k \geq 0\}$ are seen to shrink stochastically towards the origin with increasing k, as do their means and variances, and similarly for $\{\widehat{\boldsymbol{\beta}}_{Rk}; k \geq 0\}$ and $\{\widehat{\boldsymbol{\beta}}_{Sk}; k \geq 0\}$. Moreover, for each $k > 0$, $\widehat{\boldsymbol{\theta}}_{Sk}$ achieves lesser shrinkage, both in expectation and variance, than $\widehat{\boldsymbol{\theta}}_{Rk}$. Since the relations $\widehat{\boldsymbol{\beta}} = \boldsymbol{Q}\widehat{\boldsymbol{\theta}}$, $\mathrm{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{Q}\mathrm{E}(\widehat{\boldsymbol{\theta}})$, and $\mathrm{V}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{Q}\mathrm{V}(\widehat{\boldsymbol{\theta}})\boldsymbol{Q}'$ apply for all three estimators, moment properties of $\{\widehat{\boldsymbol{\beta}}_L, \widehat{\boldsymbol{\beta}}_{Rk}, \widehat{\boldsymbol{\beta}}_{Sk}\}$ follow directly. Moreover, the conditioning of $\{\widehat{\boldsymbol{\beta}}_L, \widehat{\boldsymbol{\beta}}_{Rk}, \widehat{\boldsymbol{\beta}}_{Sk}\}$, considered as linear data transformations subject to disturbances in $\boldsymbol{X}$, and of their dispersion matrices $\{\mathrm{V}(\widehat{\boldsymbol{\beta}}_L), \mathrm{V}(\widehat{\boldsymbol{\beta}}_{Rk}), \mathrm{V}(\widehat{\boldsymbol{\beta}}_{Sk})\}$, are established directly from those of $\{\widehat{\boldsymbol{\theta}}_L, \widehat{\boldsymbol{\theta}}_{Rk}, \widehat{\boldsymbol{\theta}}_{Sk}\}$, from the orthogonality of $\boldsymbol{Q}$ and invariance of condition numbers. These facts will see repeated use with little further comment.

For example, with $\boldsymbol{D}_\xi^k = \mathrm{Diag}(\xi_1/(\xi_1^2 + k), \ldots, \xi_p/(\xi_p^2 + k))$, it follows from $\widehat{\boldsymbol{\theta}}_{Rk} = \boldsymbol{D}_\xi^k\boldsymbol{U}$ that its condition number gauges sensitivity of the solution $\widehat{\boldsymbol{\theta}}_{Rk}$, considered as the function $\widehat{\boldsymbol{\theta}}_{Rk}(\boldsymbol{D}_\xi^k)$ of $\boldsymbol{D}_\xi^k$, to disturbances in the singular values of $\boldsymbol{X}$, and thus of $\widehat{\boldsymbol{\beta}}_{Rk} = \boldsymbol{Q}\widehat{\boldsymbol{\theta}}_{Rk}$ to disturbances in $\boldsymbol{X}$. These and further condition numbers are given in Table 5 for $\{\widehat{\boldsymbol{\theta}}_L, \widehat{\boldsymbol{\theta}}_{Rk}, \widehat{\boldsymbol{\theta}}_{Sk}\}$. Expressions defining $\widehat{\boldsymbol{\theta}}(\boldsymbol{D})$, as linear in $\boldsymbol{U}$ with coefficients $\boldsymbol{D}$, are given in the second column of Table 4. These in turn give condition numbers $c_1(\boldsymbol{D})$ gauging the stability of the solutions to disturbances in $\boldsymbol{X}$ through its singular values, as reported in the second column of Table 5. Further condition numbers, for $\mathrm{E}(\widehat{\boldsymbol{\theta}}) = T(\boldsymbol{\theta})$ as a parameter transformation, and for the dispersion matrix $\mathrm{V}(\widehat{\boldsymbol{\theta}})$, occupy the third and fourth columns of Table 5. Moreover, entries in Table 5 follow directly from Table 4 and the definition of $c_1(\cdot)$, on recalling that elements of $\boldsymbol{D}_\xi$ are ordered as $\{\xi_1 \geq \cdots \geq \xi_p > 0\}$. Moreover, equivalent labels for the rows of Table 5 are $\{\widehat{\boldsymbol{\beta}}_L, \widehat{\boldsymbol{\beta}}_{Rk}, \widehat{\boldsymbol{\beta}}_{Sk}\}$, and for its columns are $\{c_1[\widehat{\boldsymbol{\beta}}(\boldsymbol{X})], c_1[T(\boldsymbol{\beta})], c_1[\mathrm{V}(\widehat{\boldsymbol{\beta}})]\}$, for reasons cited.

Note that condition numbers $c_1[\widehat{\boldsymbol{\beta}}_{Sk}(\boldsymbol{X}_k)] = c_1(\boldsymbol{X}_k)$, and $c_1[\mathrm{V}(\widehat{\boldsymbol{\beta}}_{Sk})] = c_1(\boldsymbol{X}_k'\boldsymbol{X}_k)$, parallel those for $\widehat{\boldsymbol{\beta}}_L$, as both are *OLS* in their respective models. Moreover, the condition number $c_1[\widehat{\boldsymbol{\beta}}_{Sk}(\boldsymbol{X}_k)] = (\xi_1^2 + k)^{\frac{1}{2}}/(\xi_p^2 + k)^{\frac{1}{2}}$, and its square $c_1[\mathrm{V}(\widehat{\boldsymbol{\beta}}_{Sk})]$, both decrease *monotonically* with increasing k, thus assuring incremental improvement in conditioning for surrogate

Table 5
Condition numbers $c_1(\boldsymbol{D})$ for data transformations $\widehat{\boldsymbol{\theta}}(\boldsymbol{D}) = \boldsymbol{DU}$, for parameter transformations $\mathrm{E}(\widehat{\boldsymbol{\theta}}) = T(\boldsymbol{\theta})$, *and for* $\mathrm{V}(\widehat{\boldsymbol{\theta}})$, *for each of* $\{\widehat{\boldsymbol{\theta}}_L, \widehat{\boldsymbol{\theta}}_{Rk}, \widehat{\boldsymbol{\theta}}_{Sk}\}$.

Estimator	$c_1(\boldsymbol{D})$	$c_1\,[T(\boldsymbol{\theta})]$	$c_1[\mathrm{V}(\widehat{\boldsymbol{\theta}})]$
$\widehat{\boldsymbol{\theta}}_L$	$\frac{\xi_1}{\xi_p}$	1.00	$\frac{\xi_1^2}{\xi_p^2}$
$\widehat{\boldsymbol{\theta}}_{Rk}$	$\frac{\max\{\xi_i/(\xi_i^2+k)\}}{\min\{\xi_i/(\xi_i^2+k)\}}$	$\frac{\xi_1^2(\xi_p^2+k)}{\xi_p^2(\xi_1^2+k)}$	$\frac{\max\{\xi_i^2/(\xi_i^2+k)^2\}}{\min\{\xi_i^2/(\xi_i^2+k)^2\}}$
$\widehat{\boldsymbol{\theta}}_{Sk}$	$\frac{\sqrt{\xi_1^2+k}}{\sqrt{\xi_p^2+k}}$	$\frac{\xi_1\sqrt{\xi_p^2+k}}{\xi_p\sqrt{\xi_1^2+k}}$	$\frac{\xi_1^2+k}{\xi_p^2+k}$

Table 6
Condition numbers $c_1(\widehat{\boldsymbol{\beta}}_{Rk})$ and $c_1(\widehat{\boldsymbol{\beta}}_{Sk})$ for $\widehat{\boldsymbol{\beta}}_{Rk}(\boldsymbol{Z})$ and $\widehat{\boldsymbol{\beta}}_{Sk}(\boldsymbol{Z})$, as data transformations subject to disturbances in $\boldsymbol{Z}$; and for $\mathrm{V}(\widehat{\boldsymbol{\beta}}_{Rk})$ and $\mathrm{V}(\widehat{\boldsymbol{\beta}}_{Sk})$; *the maximal VIFs $V_M(\widehat{\boldsymbol{\beta}}_{Rk})$ and $V_M(\widehat{\boldsymbol{\beta}}_{Sk})$; and the Frobenius distance $D_Z(\boldsymbol{Z}_k) = \|\boldsymbol{Z} - \boldsymbol{Z}_k\|_F$, under various choices for k.*

k	$c_1(\widehat{\boldsymbol{\beta}}_{Rk})$	$c_1(\widehat{\boldsymbol{\beta}}_{Sk})$	$V_M(\widehat{\boldsymbol{\beta}}_{Rk})$	$V_M(\widehat{\boldsymbol{\beta}}_{Sk})$	$c_1[\mathrm{V}(\widehat{\boldsymbol{\beta}}_{Rk})]$	$c_1[\mathrm{V}(\widehat{\boldsymbol{\beta}}_{Sk})]$	$D_Z(\boldsymbol{Z}_k)$
0.0004	33.1584	96.1565	141.5345	1, 146.399	1, 099.4770	9, 246.064	0.0140
0.005	9.0537	28.8348	10.8874	108.0918	81.9695	831.4473	0.0654
0.010	8.1707	20.4561	9.2481	56.6915	66.7610	418.4530	0.0974
0.030	11.6724	11.8596	21.2905	21.2197	136.2440	140.6508	0.1847
0.050	15.1539	9.2114	34.0995	13.6552	229.6392	84.8507	0.2511
0.070	17.8166	7.8046	42.5990	10.2639	317.4320	60.9119	0.3083
0.090	20.4222	6.8997	51.7827	8.3166	417.0673	47.6061	0.3598
0.230	29.6720	4.3868	100.5675	3.9338	880.4276	19.2438	0.6429
0.616964	53.4183	2.7932	250.4309	2.0374	2, 853.5130	7.8022	1.1769
1.000	66.6915	2.2797	451.5788	1.5976	4, 447.7550	5.1968	1.5745

estimators. By contrast, condition numbers for $\{\widehat{\boldsymbol{\theta}}_{Rk}; k \geq 0\}$ and $\{\widehat{\boldsymbol{\beta}}_{Rk}; k \geq 0\}$ are more convoluted and will be examined next in Section 4.6.

4.6 Case Studies: Continued

Returning again with $\boldsymbol{Z}'\boldsymbol{Z}$ in correlation form, we find the Hospital Manpower Data to be exceedingly ill-conditioned: $c_1(\boldsymbol{Z}'\boldsymbol{Z}) = 77{,}754.86$; the maximal *VIF* for *OLS* is $V_1 = VIF(\widehat{\beta}_1) = 9,595.685$; and other *VIF*s are listed in Table 7 at $k = 0$. Using the data at hand, we next examine the ridge and surrogate ridge solutions with regard to issues raised in Section 4.2. Table 6 lists condition numbers and other quantities affiliated with $\{\widehat{\boldsymbol{\beta}}_{Rk}; k \geq 0\}$ and $\{\widehat{\boldsymbol{\beta}}_{Sk}; k \geq 0\}$, for values of k as listed. The table heading identifies $\widehat{\boldsymbol{\beta}}_{Rk}(\boldsymbol{Z})$ and $\widehat{\boldsymbol{\beta}}_{Sk}(\boldsymbol{Z})$ both in terms of $\boldsymbol{Z}$, so that the cited condition numbers $c_1[\widehat{\boldsymbol{\beta}}_{Rk}(\boldsymbol{Z})]$ and $c_1[\widehat{\boldsymbol{\beta}}_{Sk}(\boldsymbol{Z})]$ refer to the sensitivity of solutions to disturbances in $\boldsymbol{Z}$. To gauge effects of excessive dispersion in $\widehat{\boldsymbol{\beta}}_{Rk}$, the maximal *VIF*s, namely $V_M(\widehat{\boldsymbol{\beta}}_{Rk})$, initially decrease to 8.1707 at $k = 0.010$ and increase thereafter. Similarly, $c_1[\mathrm{V}(\widehat{\boldsymbol{\beta}}_{Rk})]$ temporarily decreases over $k \in [0, 0.015]$ to 55.4470, but then increases. By comparison, both the maximal *VIF*s for $\widehat{\boldsymbol{\beta}}_{Sk}$, and condition numbers $c_1[\mathrm{V}(\widehat{\boldsymbol{\beta}}_{Sk})]$, decrease *monotonically* with increasing k. Although initially larger, $V_M(\widehat{\boldsymbol{\beta}}_{Sk})$ approximates $V_M(\widehat{\boldsymbol{\beta}}_{Rk})$ at $k = 0.030$, whereas the ratio $V_M(\widehat{\boldsymbol{\beta}}_{Rk})/V_M(\widehat{\boldsymbol{\beta}}_{Sk})$ escalates markedly thereafter.

Stability of the solutions $\widehat{\boldsymbol{\beta}}_{Rk}$, as indexed by $c_1[\widehat{\boldsymbol{\beta}}_{Rk}(\boldsymbol{Z})]$, initially improves, achieving maximal stability at $k = 0.015$, where $c_1[\widehat{\boldsymbol{\beta}}_{Rk}(\boldsymbol{Z})] = 7.4463$ is minimal, but then erodes. In contrast, despite higher beginning values, the surrogate condition numbers $c_1[\widehat{\boldsymbol{\beta}}_{Sk}(\boldsymbol{Z})]$ decrease *monotonically*

Table 7
Variance inflation factors for $\widehat{\beta}_{Rk}$; condition numbers for correlation matrices $C(\widehat{\beta}_{Rk})$ and for $T(\beta) = \mathrm{E}(\widehat{\beta}_{Rk})$, for designated values of k.

k	*VIF1*	*VIF2*	*VIF3*	*VIF4*	*VIF5*	$c_1[C(\widehat{\beta}_{Rk})]$	$c_1\,[T(\beta)]$
0.000	9,595.68	7.9406	8931.449	23.2887	4.2794	5,4756.83	1.0000
0.0004	141.5345	7.8481	133.0221	13.0512	3.3997	576.8409	8.4095
0.005	7.1047	7.1379	7.8349	10.8874	2.9972	89.50392	93.5175
0.010	8.0001	6.4919	8.8456	9.2481	2.6830	75.66936	185.8150
0.030	19.7743	4.7268	21.2905	5.6339	2.0003	109.4703	552.8219
0.050	32.0013	3.6885	34.0995	4.0168	1.6988	177.2545	916.3722
0.070	42.5990	3.0187	45.0695	3.1473	1.5363	227.9178	1,276.515
0.090	51.7827	2.5598	54.4589	2.6269	1.4377	267.3171	1,633.297
0.230	100.5675	1.3868	102.4723	1.6364	1.2446	441.9639	4,040.511
0.616964	250.4309	1.0879	243.2535	1.8791	1.3541	1,047.931	9,965.795
1.000	451.5788	1.2184	430.5957	2.4738	1.5386	2,174.418	1,4961.96

Table 8
Variance inflation factors for $\widehat{\beta}_{Sk}$; condition numbers for correlation matrices $C(\widehat{\beta}_{Sk})$ and for $T(\beta) = \mathrm{E}(\widehat{\beta}_{Sk})$, for designated values of k.

k	*VIF1*	*VIF2*	*VIF3*	*VIF4*	*VIF5*	$c_1[C(\widehat{\beta}_{Sk})]$	$c_1\,[T(\beta)]$
0.000	9,595.68	7.9406	8,931.449	23.2887	4.2794	54,756.83	1.0000
0.0004	1,146.399	7.8846	1,068.211	14.2203	3.5089	5,091.248	2.8999
0.005	108.0918	7.5147	101.7946	12.0488	3.2099	440.5738	9.6704
0.010	56.6915	7.1607	53.8459	11.0379	3.0219	233.7461	13.6314
0.030	21.2197	6.0737	20.5412	8.4506	2.5374	93.1862	23.5122
0.050	13.6552	5.3181	13.3380	6.9511	2.2606	63.4167	30.2716
0.070	10.2639	4.7584	10.0777	5.9605	2.0792	48.8365	35.7284
0.090	8.3166	4.3258	8.1934	5.2538	1.9500	39.9124	40.4141
0.230	3.9338	2.8218	3.9092	3.1133	1.5493	17.9102	63.5650
0.616964	2.0374	1.7669	2.0330	1.8380	1.2710	7.5371	99.8288
1.000	1.5976	1.4635	1.5957	1.4981	1.1781	5.0614	122.3191

with increasing k, the trends $c_1[\widehat{\beta}_{Rk}(\boldsymbol{Z})] = 11.7723 = c_1[\widehat{\beta}_{Sk}(\boldsymbol{Z})]$ crossing at $k = 0.03045$. The Frobenius error in approximating $\boldsymbol{Z}$ by $\boldsymbol{Z}_k$ is tabulated as the final column of Table 6. Relative changes, given by $\|\boldsymbol{Z} - \boldsymbol{Z}_k\|_F / \|\boldsymbol{Z}\|_F$, are 0.1123 at $k = 0.05$, ranging up to 0.5263 at $k = 0.616964$, where the denominator takes the value $\|\boldsymbol{Z}\|_F = 2.236068$.

Finer details, in Tables 7 and 8, support several entries in Table 6. Specifically, Table 7 examines the evolution with k of *VIF*s and conditioning of correlation matrices for $\widehat{\beta}_{Rk}$. Values for $c_1[C(\widehat{\beta}_{Rk})]$ are included, as the cited inequality of Berk (1977) applies in each case. It is found that the value $c_1[C(\widehat{\beta}_{Rk})] = 61.4449$ at $k = 0.0173$ is minimal. Condition numbers for $\mathrm{E}(\widehat{\beta}_{Rk}) = T(\beta)$, as listed in the last column of Table 7, increase explosively with increasing k. Corresponding values appear in Table 8 for $\widehat{\beta}_{Sk}$, to include the evolution of *VIF*s, $c_1[C(\widehat{\beta}_{Sk})]$, and $\mathrm{E}(\widehat{\beta}_{Sk}) = T(\beta)$. Further computations show that the crossing $c_1[C(\widehat{\beta}_{Rk})] = 99.56217 = c_1[C(\widehat{\beta}_{Sk})]$ occurs at $k = 0.02750$.

In summary, the following trends may be noted from the Hospital Manpower Data.

(i) Condition numbers and maximal *VIF*s evolve erratically with k for $\{\widehat{\beta}_{Rk}; k \geq 0\}$, decreasing to a minimum and then increasing; however, these decrease monotonically for the surrogate solutions $\{\widehat{\beta}_{Sk}; k \geq 0\}$.

(ii) Ridge solutions $\{\widehat{\beta}_{Rk}; k \geq 0\}$ hold an early advantage over $\{\widehat{\beta}_{Sk}; k \geq 0\}$ in maximal *VIF*s and condition numbers, dropping away from *OLS* more quickly up to about $k = 0.030$, where $\widehat{\boldsymbol{\beta}}_{Rk}$ and $\widehat{\boldsymbol{\beta}}_{Sk}$ are comparable. Thereafter, condition numbers and maximal *VIF*s for $\widehat{\boldsymbol{\beta}}_{Rk}$ escalate, sometimes explosively, whereas corresponding values for $\widehat{\boldsymbol{\beta}}_{Sk}$ decrease monotonically.

(iii) *VIF*s for individual estimators in $\widehat{\boldsymbol{\beta}}_{Rk}$ initially decrease, then increase, but changes occur at differing values of k across the five estimators, as seen in Table 7. By contrast, for $\widehat{\boldsymbol{\beta}}_{Sk}$ Table 8 shows the maximal *VIF* to be $V_M(\widehat{\boldsymbol{\beta}}_{Sk}) = VIF(\widehat{\beta}_1)$ for all cases, independently of k.

(iv) As noted, Hoerl & Kennard (1970, p. 65), postulate that the system $\{(\boldsymbol{Z}'\boldsymbol{Z} + k\boldsymbol{I}_p)\boldsymbol{\beta} = \boldsymbol{Z}'\boldsymbol{Y};\ k \geq 0\}$, at a certain value of k, "will stabilize and have the general characteristics of an orthogonal system." However, if one takes *VIF*s and condition numbers near unity as hallmarks of an orthogonal system, their claim is not supported in the Hospital Manpower Data. On these same grounds, however, the surrogate system $\{\boldsymbol{Z}_k'\boldsymbol{Z}_k\boldsymbol{\beta} = \boldsymbol{Z}_k'\boldsymbol{Y};\ k \geq 0\}$ does appear to progress monotonically with k towards orthogonality.

5 Conclusions

A vast and expanding compendium continues to masquerade as the theory, methodology, and confirmation of ridge regression as a palliative for ill-conditioned models. Mathematical foundations long have been alleged to rest on LaGrange's constrained optimization. However, little of the collective body of ridge regression coheres logically with this claim. Contrary to conventional wisdom, we have shown (i) that solutions of given length need not be minimizing; (ii) that constraints on admissible values for k lurk within each problem; (iii) that total variance of constrained solutions is increasing in k, rather than decreasing, as claimed; (iv) that the squared bias is not necessarily increasing in k; and (v) that neither variance plus squared bias, nor the ridge traces, need be continuous in k, as assumed throughout the literature. These facts alone challenge the meaning of numerous simulation studies scattered throughout the literature, purporting to compare alternative criteria for choosing k *inter se* and with *OLS*, when such choices have ignored constraints on k. Thus aggregates of minimizing/non-minimizing solutions have been compared with other such aggregates, to the effect of total obfuscation. These anomalies trace to misuse of LaGrange's Principle; failure to identify singularities; falsely linking k one-to-one with c^2; and asserting incorrectly that a solution $\widehat{\boldsymbol{\beta}}_{Rk}$ of given length is minimizing.

Constrained minimization having failed, principles based on conditioning of linear systems are developed in Section 4. One type of conditioning traditionally has been cited in support for $\widehat{\boldsymbol{\beta}}_{Rk}$. Its limitations prompt the use of *surrogate ridge* solutions instead, to adjust for ill-conditioning of $\boldsymbol{X}$ on both sides of the *OLS* equations. Extensive numerical studies in Section 4.6 re-examine the Hospital Manpower Data in a manner complementary to Myers (1990). Citing variance inflation factors and several condition numbers as critical to the pedigree of ridge solutions $\widehat{\boldsymbol{\beta}}_{Rk}$, it is demonstrated that none stabilizes on increasing k. In contrast, for the surrogate solutions $\widehat{\boldsymbol{\beta}}_{Sk}$, these properties are uniformly enhanced as k evolves, excluding $\mathrm{E}(\widehat{\boldsymbol{\beta}}_{Sk}) = T(\boldsymbol{\beta})$. It is seen that $\widehat{\boldsymbol{\beta}}_{Rk}$ is better within a narrow range for small k, but its *VIF*s and condition numbers then become excessive. In short, ridge regression often exhibits some of the very pathologies it is intended to redress.

In retrospect, the Hoerl & Kennard (1970) criticism that *OLS* "does not have built into it a method for portraying sensitivity of the solutions to the estimation criterion," apparently prompted their turning to constrained solutions. Their ridge traces then would link such matters as stability, sign changes, and other features of the analysis as k evolves, to be linked in turn with the constraint criterion $\{\boldsymbol{\beta}'\boldsymbol{\beta} = c^2\}$. Our developments show conclusively that even these aspirations, however well intended, have failed.

References

Allen, D.M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, **16**, 125–127.

Balakrishnan, A.V. (1963). An operator theoretic formulation of a class of control problems and a steepest descent method of solution. *J. Contr.*, **1**, 109–127.

Beaton, A.D., Rubin, D. & Barone, J. (1976). The acceptability of regression solutions: Another look at computational accuracy. *J. Am. Statist. Assoc.*, **71**, 158–168.

Belsley, D.A., Kuh, E. & Welsch, R.E. (1980). *Regression Diagnostics*. New York: Wiley.

Berk, K.N. (1977). Tolerance and condition in regression computations. *J. Am. Statist. Assoc.*, **72**, 863–866.

Bunke, O. (1975). Minimax linear, ridge and shrunken estimators for linear parameters. *Math. Operationsforsch. u. Statist.*, **6**, 697–701.

Davies, R.B. & Hutton, B. (1975). The effect of errors in the independent variables in regression. *Biometrika*, **62**, 383–392.

Dempster, A.P., Schatzoff, M. & Wermuth, N. (1977). A simulation study of alternatives to ordinary least squares. *J. Am. Statist. Assoc.*, **72**, 77–106.

Draper, N.R. (1963). 'Ridge analysis' of response surfaces. *Technometrics*, **5**, 469–479.

Golub, G.H., Heath, C.G. & Wahba, G. (1979). Generalized cross validation as a method for choosing a good ridge parameter. *Technometrics*, **21**, 215–223.

Hocking, R.R. (1976). The analysis and selection of variables in linear regression. *Biometrics*, **32**, 1–49.

Hoerl, A.E. (1962). Application of ridge analysis to regression problems. *Chem. Eng. Prog.*, **58**, 54–59.

Hoerl, A.E. (1964). Ridge analysis. *Chem. Eng. Progr. Symposium Series*, **60**, 67–77.

Hoerl, A.E. & Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Hoerl, A.E., Kennard, R.W. & Baldwin, K.F. (1975). Ridge regression: Some simulations. *Commun. Stat.*, **4**, 105–123.

Horn, R.A. & Johnson, C.R. (1985). *Matrix Analysis*. Cambridge, UK: Cambridge University Press.

Levenberg, K. (1944). A method for the solution of certain non-linear problems in least squares. *Quart. Appl. Math.*, **2**, 164–168.

Mallows, C.L. (1973). Some comments on C_p. *Technometrics*, **15**, 661–675.

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

Marquardt, D.W. & Snee, R.D. (1975). Ridge regression in practice. *Am. Stat.*, **29**, 3–20.

Marshall, A.W. & Olkin, I. (1979). *Inequalities: Theory of Majorization and Its Applications*. New York: Academic Press.

Myers, R.H. (1990). *Classical and Modern Regression with Applications*, 2nd ed. Boston, MA: PWS-KENT Publishing Co.

Myoken, H. & Uchida, Y. (1977). The generalized ridge estimator and improved adjustments for regression parameters. *Metrika*, **24**, 113–124.

Nostrand, C. van (1980). Comment. *J. Am. Statist. Assoc.*, **75**, 92–94.

Riley, J. (1955). Solving systems of linear equations with a positive definite, symmetric but possibly ill–conditioned matrix. *Math. Tables Aids Comput.*, **9**, 96–101.

Stoer, J. & Witzgall, C. (1970). *Convexity and Optimization in Finite Dimensions I*. New York: Springer-Verlag.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B*, **58**, 267–288.

Zhang, R. & McDonald, G.C. (2005). Characterization of ridge trace behavior. *Commun. Stat.-Theor. Methods*, **34**, 1487–1501.

Résumé

Les erreurs persistent dans la régression ridge, ses bases, et son utilisation, comme déterminé en Hoerl et Kennard (1970) et plus tard. Il ne faut ni que les estimateurs ridge se réduisent au minimum ni qu'un paramètre ridge soit admissible. Les estimateurs conventionnels ne sont pas les solutions de Lagrange contraintes aux longueurs fixes, comme souvent prétendu, car de telles solutions sont singulières. D'une littérature vaste—sur l'évaluation, la prévision, la validation croisée, le choix du paramètre ridge, et sujets alliés, sujets collectivement connus sous le nom de régression ridge—peu est issu de la minimisation contrainte, même vis à vis les contraintes d'inégalitié. Le problème remonte à une mauvaise application du principe de Lagrange, au manque d'identifier des singularités, et aux liens mal placés entre les contraintes et les paramè tres ridge. Des principes alternatifs, basés sur des numéraux de condition, peuvent être vus comme validant ridge conventionnelle et la régression de ridge succédanée, ce dernier à être défini. Les études numériques illustrent que la régression ridge, comme practiquée, montrent souvent des pathologies qu'il vise à redresser.

Appendix

LaGrange's Principle. Given differentiable functions $f(u_1, \ldots, u_p)$ and $g(u_1, \ldots, u_p)$ such that the gradient $\nabla\, g(u_1, \ldots, u_p) \neq \mathbf{0}$ on $G_0 = \{\boldsymbol{u} \in \mathbb{R}^p : g(\boldsymbol{u}) = 0\}$, the problem is to minimize $f(u_1, \ldots, u_p)$ subject to the constraint $g(u_1, \ldots, u_p) = 0$. Define the Lagrangian $L(u_1, \ldots, u_p, \lambda) = f(u_1, \ldots, u_p) + \lambda[g(u_1, \ldots, u_p) - 0]$, and consider the equations

$$\nabla f(u_1, \ldots, u_p) = \lambda \nabla g(u_1, \ldots, u_p) \tag{A.1}$$

$$\partial L(u_1, \ldots, u_p, \lambda)/\partial\lambda = [g(u_1, \ldots, u_p) - 0]. \tag{A.2}$$

It is necessary that gradient vectors be parallel, i.e. (A.1), whereas (A.2) recovers the given constraint. LaGrange's Principle requires solving these equations in the $k + 1$ unknowns $\{u_1, \ldots, u_p, \lambda\}$. If minimizing $f(u_1, \ldots, u_p)$ subject to inequality constraints $g(u_1, \ldots, u_p) \leq 0$, necessary conditions for $\boldsymbol{u}^*$ to be a solution are $C_1\colon g(\boldsymbol{u}^*) \leq 0$; $C_2 : \nabla_{\boldsymbol{u}}\, L(\boldsymbol{u}^*, \lambda^*) = \mathbf{0}$; $C_3 : \lambda^*\, g(\boldsymbol{u}^*) = \mathbf{0}$; and $C_4 : \lambda^* \geq 0$. For reference see Stoer & Witzgall (1970), for example. It remains to determine that solutions thus obtained are minimizing.

For constrained least squares, the Lagrangian is now $L(\beta_1, \ldots, \beta_p, \lambda) = Q(\boldsymbol{\beta}) + \lambda(\boldsymbol{\beta}'\boldsymbol{\beta} - c^2)$ with $Q(\boldsymbol{\beta})$ as before, where

$$(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}_p)\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y} \tag{A.3}$$

$$\boldsymbol{\beta}'\boldsymbol{\beta} = c^2 \tag{A.4}$$

are to be solved for the $k + 1$ unknowns $(\beta_1, \ldots, \beta_p, \lambda)$. Designate these as $\{\widehat{\boldsymbol{\beta}}_c, \lambda_c\}$ such that $\{\widehat{\boldsymbol{\beta}}'_c\widehat{\boldsymbol{\beta}}_c = c^2\}$, as apparently intended by Hoerl & Kennard (1970). Conditions corresponding to C_1–C_4 apply under inequality constraints $\{\boldsymbol{\beta}'\boldsymbol{\beta} \leq c^2\}$. Again it remains to determine that solutions are minimizing.

[*Received November 2006, accepted December 2007*]