McDonald, G.C. (1982). Ridge estimators as constrained generalized least squares. In *Statistical Decision Theory and Related Topics III*, Eds. S.S. Gupta & J.O. Berger, vol. 2, pp. 183–191. New York, NY: Academic Press.

Meeter, D.A. (1966). On a theorem used in nonlinear least squares. *SIAM J. Appl. Math.*, **14**, 1176–1179.

Myers, R.H. (1989). *Classical and Modern Regression With Applications*, 2nd ed. Boston, MA: PWS-KENT Publishing Co.

Sardy, S. (2008). On the practices of rescaling covariates. *Internat. Statist. Rev.*, **76**, 285–297.

Smith, G. & Campbell, F. (1980). A critique of ridge regression methods. *J. Amer. Statist. Assoc.*, **75**, 74–81.

Sundaram, R.K. (1996). *A First Course in Optimization Theory*, 2nd ed. Cambridge: Cambridge University Press.

Résumé

Plusieurs anomalies ont été récemment relevées par Jensen et Ramirez (2008) dans les fondements théoriques de la "ridge regression" considérée dans une perspective de moindres carrés constraints. Certaines de ces anomalies ont été attribuées au comportement non monotone de la norme des "ridge-estimateurs" non contraints, ainsi qu'au caractère non suffisant du principe de Lagrange. Nous indiquons dans cet article que, pour une valeur fixée de **Y**, la norme des ridge-estimateurs correspondant à une base donnée sont strictement monotones. En outre, les conditions assurant le caractère suffisant du principe de Lagrange sont satisfaites pour un ensemble adéquat de valeurs du paramètre contraint. L'origine des anomalies relevées se trouve donc ailleurs. Cette apparente contradiction prend son origine, dans le contexte de l'étude d'un ensemble de données particulier, dans la confusion entre les estimateurs du vecteur de paramètres β correspondant à différentes paramétrisations (associées à différents choix d'une base) et/ou à différentes normes. Afin d'éviter ce type de confusion, il est suggéré d'indexer le paramètre de façon adéquate au moyen de la base choisie.

Anomalies in Ridge Regression: Rejoinder

Donald R. Jensen and Donald E. Ramirez

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

1 Introduction

We wish to thank Professors Kapat and Goel (2010) (hereafter KG(2010)) for clarifying the non-invariance of ridge regression to the choice of parametrization, a concern voiced by Smith and Campbell (1980), their discussants, and others. KG(2010) focus mainly on one aspect of our "Foundations" paper (hereafter JR(2008)), namely, constrained optimization in linear inference. We mostly concede their corrections, noting that essentials are found in Marquardt (1963), Meeter (1966), Hoerl & Kennard (1970), and recently Davidov (2006). However, the principal anomaly of JR(2008) remains: The theory and practice of ridge regression do *not* rest on constrained optimization.

Here we trace conventions for taking $\{Y = X\beta + \epsilon\}$, in original coordinates, through centering and scaling into $\{Y = Z\theta + \epsilon\}$ with $A = Z'Z$ in "correlation form", the model M_R of KG(2010), but now distinguishing θ from β. In practice, the original scale typically is of overriding concern to the analyst; hence the routine mapping back to that scale. In fact, an algorithm in section 4.2 of JR(2008) shows connections between $\{Y = Z\theta + \epsilon\}$, where solutions diminish monotonically in length, and lengths of projections back onto the original scale, which need not be monotone.

2 Standardization

Constrained optimization in regression traces to Marquardt (1963), seeking a compromise between Gauss–Newton and gradient methods in non-linear estimation, and solving iteratively through locally linear models. Noting that gradient methods are not scale invariant, Marquardt (1963, p. 436) for convenience scales the β-space using standard deviations of the columns of X, transforming elements of $X'X$ into "simple correlation coefficients", as used widely in least-squares problems "for improving the numerical aspects of computing procedures". As in Belsley (1986), the latter includes the conditioning of linear systems and stability of their solutions. This standardization pervades much of the ridge regression literature, beginning with Hoerl & Kennard (1970) and onward.

That solutions are mapped routinely back onto the original scale is seen in the *Ridge* option of *Proc Reg* in the *SAS* system, as used in Myers (1990) for the Hospital Manpower Data and as verified in JR(2008). An antecedent is Marquardt's (1963) algorithm: At each iteration the local linear model is transformed into "correlation form"; the constrained solution is found; and this is projected back onto the original scale before proceeding to the next iterate. Letting $D_s = \text{Diag}(s_1^{-1}, \ldots, s_p^{-1})$ as in KG(2010), the constraint $\{\theta'\theta = c^2\}$ maps back to $\{\beta' D_s^{-2}\beta = c^2\}$, often counter-intuitive to the user and subject to the vagaries of typically irrelevant lengths of the columns of X. That ridge solutions $\|\widehat{\theta}_{R_k}\|^2 = \widehat{\beta}'_{R_k} D_s^{-2}\widehat{\beta}_{R_k}$ decrease monotonically in k is an artifact of an algorithm chosen for numerical stability. To the contrary, users instead may insist that $\|\widehat{\beta}_{R_k}\|$ is essential, and then seek the value $\widehat{\beta}_R(\widehat{k})$ so as (i) to achieve a natural constraint $\{\|\widehat{\beta}_{R_k}\| = c\}$, and (ii) to minimize the residual sum of squares on that scale. To these ends, the algorithm in section 4.2 of JR(2008) bridges this gap, demonstrating explicitly the manner in which solutions from $\{Y = Z\theta + \epsilon\}$, when projected as lengths onto the original scale, will be minimizing there. Specifically, for fixed c again define the set

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

and let $k_c = \min\{\Lambda(c)\}$. Then $\widehat{\beta}_R(k_c)$ achieves its designated constraint and is minimizing, all on the original scale. KG(2010) claim that $\Lambda(c)$ is either a singleton set or is empty. While true for solutions in "correlation form", the cardinality of $\Lambda(c)$ may increase when projected back onto the original scale, as noted and illustrated correctly in JR(2008).

3 Conclusions

Pretensions to the contrary, ridge regression remains ill-posed mathematically. Under objective constraints, either $\{\beta'\beta = c^2\}$ or $\{\beta'\beta \leq c^2\}$ in a generic model $\{Y = X\beta + \epsilon\}$, constrained optimization returns a solution $[\widehat{\beta}_R(\widehat{k}), \widehat{k}, c^2]$. That an objective c^2 typically is missing, is acknowledged in the myriad choices for k advocated in the literature, including the early "ridge trace" of Hoerl & Kennard (1970).

The overriding determinant of the need for ridge regression, or other biased estimation, is conditioning of the system $\{X'X\beta = X'Y\}$. Possibilities to improve the conditioning of $X'X$

include $\{X'X \rightarrow (X'X + B)\}$ with B positive definite. For $k > 0$ it is known (Riley, 1955) that $(X'X + kI_p)$ is better conditioned than $X'X$, cited as justification for ridge regression in Marshall & Olkin (1970). Nonetheless, constrained optimization is not a panacea: Given any "ridge-type" system $\{(X'X + k\Delta)\beta = X'Y\}$, under generalized constraints as in section 1 of KG(2010) and McDonald (1982), it remains open (apart from $\Delta = kI_p$) as to circumstances where its conditioning might be enhanced over the original. If not, this could further undermine conditioning of the system, and exacerbate the problem through an inappropriate and counter-productive choice of constrained optimization.

The focus of KG(2010) on constrained optimization, and our continuation in this reply, are largely beside the point, distracting, and obscuring from the main thesis of JR(2008): Whatever its niceties, constrained optimization has essentially nothing to do with ridge regression as practiced. Numerous procedures have been advanced for determining k, to include ridge traces, choices assuring dominance in mean square over Ordinary Least Squares (*OLS*) or, *inter se,* among other biased estimators; and other approaches. In particular, variances, biases, and mean square errors, as given in section 4 of Hoerl & Kennard (1970); and numerous other properties of ridge solutions adopted by generations hence: These all assert incorrectly that $\widehat{\beta}_{R_k}$, intended as a constrained estimator, is linear in the *OLS* solution $\widehat{\beta}_L$. Granted, the ridge system $\{(X'X + kI_p)\widehat{\beta}_{R_k} = X'Y\}$ does derive from initial steps of constrained optimization. However, this in itself fails to justify repeated claims that ridge regression rests on constrained optimization. To be correct, the final step must deliver the constrained solution. As noted earlier, conditioning arguments alone deliver the identical system $\{(X'X + kI_p)\widehat{\beta}_{R_k} = X'Y\}$ without recourse to constraints. In short, once again, estimators constrained to the c-sphere, or to the c-ball in Euclidean p-space, cannot be linear in $\widehat{\beta}_L$, as we emphasized from several perspectives in JR(2008). Specifically, a solution constrained to a ball of radius c *could* be unbiased, depending on how its probability is concentrated there. Nothing offered by KG(2010) can alter these facts, totally debilitating as they are to claims that constrained optimization is the foundation for ridge regression as known and practiced.

Ridge regression having failed its vaunted credentials, we introduced in JR(2008) *surrogate estimators* as modifications of ridge based on conditioning, and we showed numerically in our case study that, in contrast to ridge, surrogate estimators possess desirable monotone properties. As the ridge parameter k evolves, we since have shown in theory that ridge estimators typically exhibit erratic divergence from those of orthogonal systems, often reverting back to *OLS* in the limit. In contrast, surrogate solutions are seen to converge monotonically to those from orthogonal systems. This work appears as Jensen, D.R. & Ramirez, D.E. (2010). Surrogate models in ill-conditioned systems. *J. Statist. Plann. Inference*, **140**, 2069–2077.

References

Belsley, D.A. (1986). Centering, the constant, first-differencing, and assessing conditioning. In *Model Reliability* (Chapter 5), Eds. D.A. Belsley & E. Kuh, pp. 117–152. Cambridge, MA: MIT Press.
Note: References cited in the text but not listed in the references list are found in JR(2008) and KG(2010).