Limitations of the Least Squares Estimators; A Teaching Perspective

Diarmuid O'Driscoll
Head, Department of Mathematics and Computer Studies
Mary Immaculate College
Ireland

Donald E. Ramirez
Department of Mathematics
University of Virginia
USA

Abstract

The standard linear regression model can be written as $\mathbf{Y} = \mathbf{X}\beta+\varepsilon$ with $\mathbf{X}$ a full rank $n \times p$ matrix and $L(\varepsilon) = N(\mathbf{0}, \sigma^2\mathbf{I}_n)$. The least squares estimator is $\widehat{\beta}_L = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{XY}$ with variance-covariance matrix $Cov(\widehat{\beta}_L) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, where $Var(\varepsilon_i) = \sigma^2$. The diagonal terms of the matrix $Cov(\widehat{\beta}_L)$ are the variances of the Least Squares estimators $\widehat{\beta}_i$, $0 \leq i \leq p-1$ and the Gauss-Markov Theorem states $\widehat{\beta}_L$ is the best linear unbiased estimator. However, the OLS solutions require that $(\mathbf{X}'\mathbf{X})^{-1}$ be accurately computed and ill conditioning can lead to very unstable solutions. Tikhonov, A.N. (1943) first introduced the idea of regularisation to solve ill-posed problems by introducing additional information which constrains (bounds) the solutions. Specifically, Hoerl, A.E. (1959) added the constraint term to the least squares problem as follows: minimize $||Y - X\beta||^2$ subject to the constraint $||\beta||^2 = r^2$ for fixed r and dubbed this procedure as ridge regression. This paper gives a brief overview of ridge regression and examines the performance of three different types of ridge estimators; namely the ridge estimators of Hoerl, A.E. (1959), the surrogate estimators of Jensen, D.R. and Ramirez, D.E. (2008) and the raise estimators of Garcia, C.B., Garcia, J. and Soto, J. (2011).

Introduction

The standard linear regression model $\mathbf{Y} = \mathbf{X}\beta+\varepsilon$ has uncorrelated, zero-mean, and homoscedastic errors ε. In this paper we assume that $\mathbf{X}$ is a full rank $n\times p$ matrix containing the explanatory variables and the response vector $\mathbf{y}$ is $n \times 1$ consisting of the observed data. The Ordinary Least Squares *OLS* estimator $\widehat{\beta}_L$ is the solution of

$$\mathbf{X}'\mathbf{X}\widehat{\beta} = \mathbf{X}'\mathbf{Y} \tag{1}$$

and the Gauss-Markov Theorem states that $\widehat{\beta}_L$ is the best linear unbiased estimator. However, with economic or medical data the predictor variables may have a high level of collinearity and hence $(\mathbf{X}'\mathbf{X})^{-1}$ will be numerically difficult to calculate resulting in very unstable solutions. Small changes in the data may lead to large changes to the regression coefficients.

For example, if a repeated experiment produces the following design matrices $\mathbf{X_1}$ and $\mathbf{X_2}$ with associated $\mathbf{Y_1}$ and $\mathbf{Y_2}$

$$\mathbf{X_1} = \begin{bmatrix} 1 & 6 & 14 \\ 1 & 6 & 17 \\ 1 & 7 & 17 \\ 1 & 7 & 17 \\ 1 & 9 & 21 \end{bmatrix} \quad \mathbf{Y_1} = \begin{bmatrix} 66 \\ 74 \\ 75 \\ 73 \\ 93 \end{bmatrix}$$

$$\mathbf{X_2} = \begin{bmatrix} 1 & 6 & 14 \\ 1 & 6 & 16 \\ 1 & 7 & 17 \\ 1 & 8 & 18 \\ 1 & 9 & 22 \end{bmatrix} \quad \mathbf{Y_2} = \begin{bmatrix} 65 \\ 72 \\ 76 \\ 75 \\ 92 \end{bmatrix}$$

the least squares estimators are respectively

$$\beta_1 = \begin{bmatrix} 9.84 \\ 2.34 \\ 2.91 \end{bmatrix} \quad \beta_2 = \begin{bmatrix} 18.61 \\ -2.84 \\ 4.47 \end{bmatrix}.$$

The high condition numbers 22133 and 16067 of $\mathbf{X_1}'\mathbf{X_1}$ and $\mathbf{X_2}'\mathbf{X_2}$ respectively result in the least squares solutions being unstable.

The basic idea behind ridge regression is to trade off some bias in the estimators to gain a reduction in the variance of these estimators. Hoerl, A.E. (1959) added the penalty term to the least squares problem as follows:

$$\text{Minimize } ||\mathbf{Y} - \mathbf{X}\beta||^2 \text{ subject to } ||\beta||^2 = r^2$$

,

which is solved with Lagrange Multipliers by setting the derivatives $\frac{\partial}{\partial \beta}$ and $\frac{\partial}{\partial \lambda}$ equal to zero in the (ridge regression) **loss function**

$$\mathbf{L}(\beta, \lambda) = ||\mathbf{Y} - \mathbf{X}||\beta^2 + \lambda(||\beta||^2 - r^2).$$

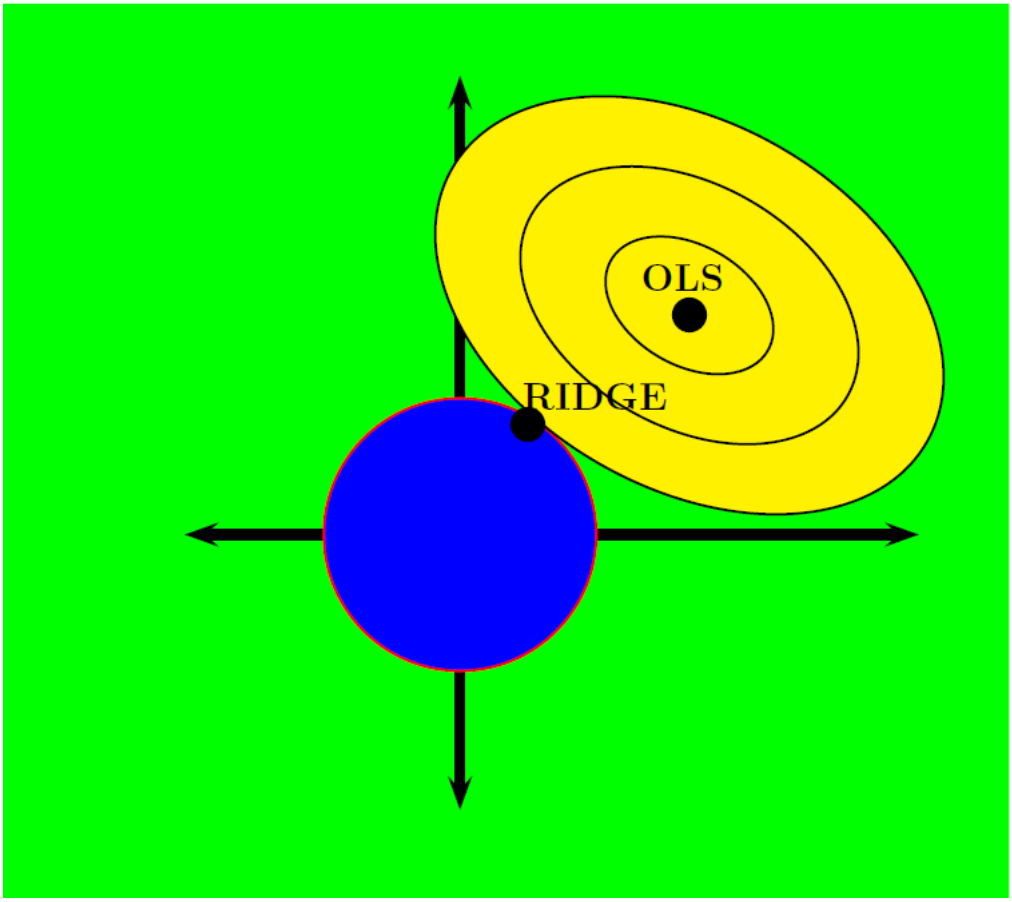
The first portion of the loss function, $||\mathbf{Y} - \mathbf{X}\beta||^2$, is the same as the ordinal least squares RSS and is referred to as the **data fidelity term** and $\lambda(||\beta||^2 - r^2)$ is referred to as the **regularization (penalty) parameter**.

Davidov (2006) proved that an equivalent problem is to minimize

$$||\mathbf{Y} - \mathbf{X}\beta||^2 \text{ subject to the constraint } ||\beta||^2 \leq r^2 \text{ with } r^2 \text{ fixed.}$$

The constraint is the convex ball in Figure 1 and the problem is a constrained optimization problem which uses quadratic programming.

Figure 1. *Geometric View of Ridge Regression*



The ellipses correspond to level curves of the residual sum of the squares (RSS) and are minimized at the ordinal least squares estimate (OLS). The penalty parameter in this case is restricting the ridge estimate to the disc. The Lagrange method gives a solution at the tangent point to the ellipse and the circle and is the trade off between the bias and the variance of the estimators and will be discussed in Section 2. Section 3 will show how to evaluate the ridge estimators of Hoerl A. E., the surrogate estimators of Jensen, D.R. and Ramirez, D.E. and the raise estimators of Garcia, C.B., Garcia, J. and Soto, J. Section 4 will summarise the properties and the results of the three estimators.

Ridge Regression

The loss function

$$\begin{aligned}\mathbf{L}(\beta,\lambda) &= \|\mathbf{Y}-\mathbf{X}\beta\|^2+\lambda(\|\beta\|^2-r^2)\\ &= <\mathbf{Y}',\mathbf{Y}>-2<\mathbf{Y},\mathbf{X}\beta>+<\mathbf{X}\beta,\mathbf{X}\beta>+\lambda<\beta',\beta>-\lambda r^2\end{aligned}$$

is minimized when

$$\frac{\partial \mathbf{L}(\beta,\lambda)}{\partial\beta} = -2\mathbf{X}'\mathbf{Y}+2\mathbf{X}'\mathbf{X}\beta+2\lambda\beta=\mathbf{0}$$

yielding the ridge estimator solution

$$\widehat{\beta}_R(r) = (\mathbf{X}'\mathbf{X}+\lambda\mathbf{I_p})^{-1}\mathbf{X}'\mathbf{Y} \qquad (2)$$

with $\widehat{\beta}_R(r)$ satisfying

$$\frac{\partial \mathbf{L}(\beta,\lambda)}{\partial\lambda} = \|\beta\|^2-r^2=0.$$

In particular, for orthogonal covariates, $\mathbf{X'X} = n\mathbf{I_p}$ and

$$\widehat{\beta}_R(r) = (n\mathbf{I_p}+\lambda\mathbf{I_p})^{-1}\mathbf{X}'\mathbf{Y}=\frac{n}{n+\lambda}\widehat{\beta}_L.$$

$$\text{As } \lambda\to 0,\ \widehat{\beta}_R(r)\to\widehat{\beta}_L \text{ and as } \lambda\to\infty,\ \widehat{\beta}_R(r)\to 0.$$

Also $\mathbf{L}(\beta, \lambda)$ is strictly convex since the Hessian

$$\frac{\partial^2 \mathbf{L}(\beta, \lambda)}{\partial \beta^2} = 2(\mathbf{X}'\mathbf{X} + \lambda \mathbf{I_p})$$

is positive definite, which guarantees that the ridge solution is unique and yields a minimum value for **L**(β, λ).

With **A** = (**X´X** + λ**I**$_\mathbf{p}$) and writing $\widehat{\beta}_R(r)$ as $\widehat{\beta}_R$

$$\begin{aligned} E(\widehat{\beta}_R) &= \mathbf{A}^{-1}\mathbf{X}'\mathbf{X}\beta \\ &= \mathbf{A}^{-1}(\mathbf{A} - \lambda \mathbf{I_p})\beta \\ &= (\mathbf{I}_p - \lambda \mathbf{A}^{-1})\beta \\ &= \beta - \lambda \mathbf{A^{-1}}\beta. \end{aligned} \tag{3}$$

$$\begin{aligned} cov(\widehat{\beta}_R) &= cov(\mathbf{A}^{-1}\mathbf{X}'\mathbf{Y}) \\ &= \mathbf{A}^{-1}\mathbf{X}' cov(\mathbf{Y})(\mathbf{A}^{-1}\mathbf{X}')' \\ &= \sigma^2 \mathbf{A}^{-1}\mathbf{X}'\mathbf{X}\mathbf{A}^{-1}. \end{aligned} \tag{4}$$

$$\begin{aligned} MSE(\widehat{\beta}_R) &= E(\|\widehat{\beta}_R - \beta\|^2) \\ &= E(\|\widehat{\beta}_R - E(\widehat{\beta}_R)\|^2) + \|E(\widehat{\beta}_R) - \beta\|^2 \\ &= tr(cov(\widehat{\beta}_R)) + \|bias(\widehat{\beta}_R)\|^2. \end{aligned} \tag{5}$$

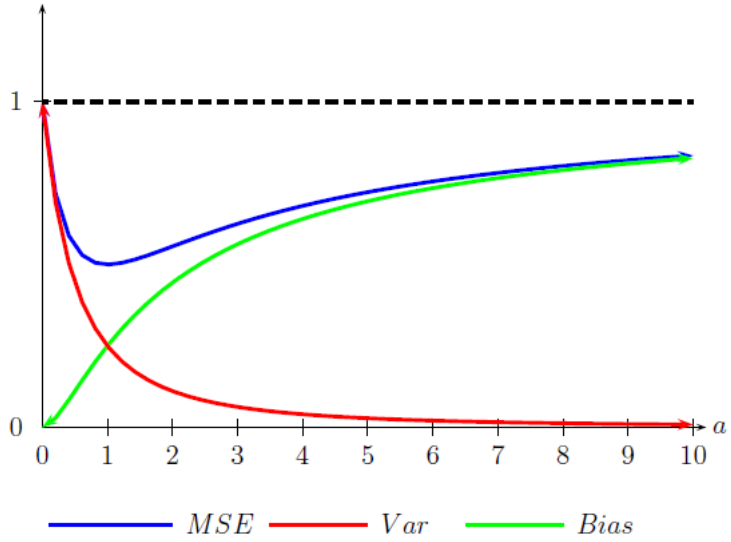
We illustrate the **trade off** between bias and variance by the following example.

If $\widehat{\beta}_L$ is $N(1, 1)$ and $\widehat{\beta}_R = \frac{1}{a+1}\widehat{\beta}_L$ $\alpha \geq 0$, then

$$\begin{aligned} MSE(\widehat{\beta}_R) &= var(\widehat{\beta}_R) + (bias(\widehat{\beta}_R))^2 \\ &= \frac{1}{(a+1)^2} + \left(\frac{1}{a+1} - 1\right)^2. \end{aligned}$$

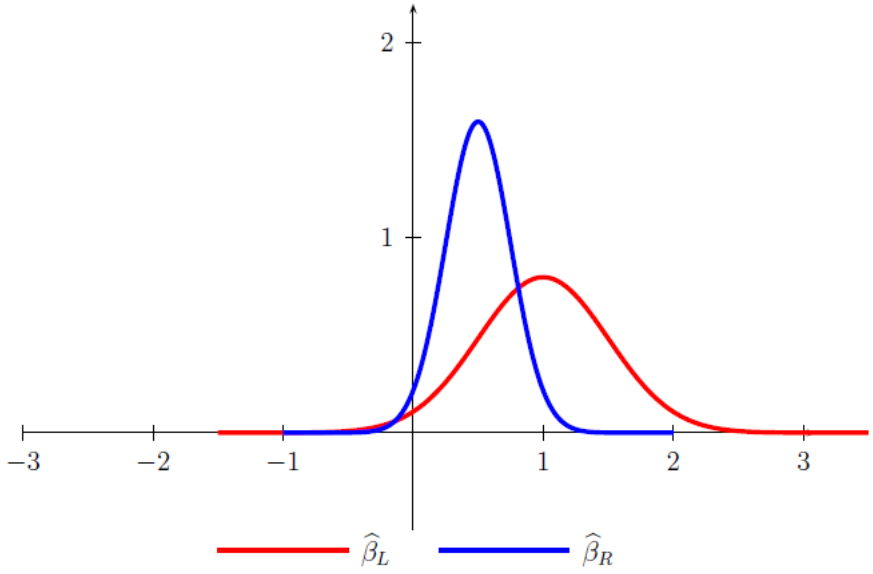
The graphs of $MSE(\widehat{\beta}_R)$; $Var(\widehat{\beta}_R)$ and $Bias(\widehat{\beta}_R)$ are illustrated in Figure 2.

Figure 2. $MSE(\widehat{\beta}_R)$; $Var(\widehat{\beta}_R)$; $Bias(\widehat{\beta}_R)$



For the optimal choice of $\alpha = 1$, $MSE(\widehat{\beta}_R) < MSE(\widehat{\beta}_L)$. In this case, the distributions of $\widehat{\beta}_L$ and $\widehat{\beta}_R$ are illustrated in Figure 3.

Figure 3. *Density functions f or* $\widehat{\beta}_L$ *and* $\widehat{\beta}_R$



Hoerl and Kennard (1970) proved that for increasing **λ**, $tr(cov(\widehat{\beta}_R))$ is a strictly decreasing function of **λ**, $\|bias(\widehat{\beta}_R)\|^2$ is a strictly increasing function of **λ** and that there always exists **λ** such that $MSE(\widehat{\beta}_R) < MSE(\widehat{\beta}_L)$.

If **D** is the matrix of eigenvalues $d_1 \geq d_2 ... \geq d_p > 0$ of **X′X**, then the eigenvalues of **A** are $(d_i + \lambda)$, $1 \leq i \leq p$. If **P** is the orthogonal transformation such that **X′X** = **PDP′** then **A** = **P**(**D**+ $\lambda \mathbf{I_p}$)**P′**. Writing α= **P′**β, the canonical variable,

$$\begin{aligned} B(\lambda) = \|bias(\widehat{\beta}_R)\|^2 &= (-\lambda \mathbf{A^{-1}}\beta)'(-\lambda \mathbf{A^{-1}}\beta) \\ &= \lambda^2 \beta'(\mathbf{A^{-1}})^2\beta \\ &= \lambda^2 \beta' \mathbf{P}(\mathbf{D}+\lambda \mathbf{I_p})^{-2}\mathbf{P}'\beta \\ &= \lambda^2 \alpha'(\mathbf{D}+\lambda \mathbf{I_p})^{-2}\alpha \\ &= \lambda^2 \sum_{i=1}^{p} \frac{\alpha_i^2}{(d_i+\lambda)^2} \end{aligned} \tag{6}$$

From Eq(7),

$$B(0) = 0, \quad \lim_{\lambda \to \infty} B(\lambda) = \sum_{i=1}^{p} \alpha_i^2$$

And *B*(**λ**) is monotone increasing as

$$\frac{dB}{d\lambda} = 2\lambda \sum_{i=1}^{p} \frac{\alpha_i^2 d_i}{(d_i+\lambda)^3} > 0, \quad \lambda > 0. \tag{7}$$

Similarly

$$\begin{aligned} V(\lambda) = tr(cov(\widehat{\beta}_R)) &= tr(\sigma^2 \mathbf{A}^{-1}\mathbf{X}'\mathbf{X}\mathbf{A}^{-1}) \\ &= \sigma^2 \sum_{i=1}^{p} \frac{d_i}{(d_i+\lambda)^2} \end{aligned} \tag{8}$$

From Eq(9),

$$V(0) = \sigma^2 \sum_{i=1}^{p} \frac{1}{d_i}, \quad \lim_{\lambda \to \infty} V(\lambda) = 0$$

And *V*(**λ**) is monotone decreasing as

$$\frac{dV}{d\lambda} = -2\sigma^2 \sum_{i=1}^{p} \frac{d_i}{(d_i+\lambda)^3} < 0, \quad \lambda > 0. \tag{9}$$

From Eq(8) and Eq(10)

$$\frac{dMSE(\widehat{\beta}_R)}{d\lambda} = 2\sum_{i=1}^{p} \frac{d_i(\lambda\alpha_i^2 - \sigma^2)}{(d_i+\lambda)^3} < 0, \quad \lambda < \sigma^2/\alpha_{max}^2$$

And hence there always exists a **λ** such that $MSE(\widehat{\beta}_R) < MSE(\widehat{\beta}_L)$

However as $\alpha_{min}^2 > 0$ and

$$\frac{dMSE(\widehat{\beta}_R)}{d\lambda} = 2\sum_{i=1}^{p} \frac{d_i(\lambda\alpha_i^2 - \sigma^2)}{(d_i+\lambda)^3} > 0, \quad \lambda > \sigma^2/\alpha_{min}^2,$$

$MSE(\widehat{\beta}_R)$ is a monotone increasing function for such **λ**. In Table 1 of Jensen and Ramirez (2010), the authors give an example of **λ** for which $MSE(\widehat{\beta}_R) > MSE(\widehat{\beta}_L)$.

Ridge, Surrogate and Raise Estimators

In this section we will discuss three standard remedies for addressing collinearity in linear regression; namely (1) the *ridge system* {($\mathbf{X}'\mathbf{X}$ + $\boldsymbol{\lambda}\mathbf{I}_p$)$\beta$ = $\mathbf{X}'\mathbf{X}$; $\boldsymbol{\lambda}\geq 0$} (Hoerl and Kennard, 1970) with solutions $\{\widehat{\beta}_R(\lambda); \lambda \geq 0\}$; (2) the *surrogate system* {($\mathbf{X}'\mathbf{X}$ + $\boldsymbol{\lambda}\mathbf{I}_p$) β = ($\mathbf{X}'_{\boldsymbol{\lambda}}\mathbf{X}_{\boldsymbol{\lambda}}$)$\beta$ =$\mathbf{X}'_{\boldsymbol{\lambda}}$Y; $\boldsymbol{\lambda}\geq 0$} (Jensen and Ramirez, 2008) with solutions $\{\widehat{\beta}_S(\lambda); \lambda \geq 0\}$ and (3) the *raise system* with solutions $\{\widehat{\beta}_{RA}(\lambda); \lambda \geq 0\}$ (Garcia *et al*, 2011).

The ridge estimators come from modifying $\mathbf{X}'\mathbf{X} \rightarrow \mathbf{X}'\mathbf{X} + \boldsymbol{\lambda}\mathbf{I}_p$ on the left side of Eq.(1) while the Jensen and Ramirez surrogate estimators modify the design matrix $\mathbf{X} \rightarrow \mathbf{X}_{\boldsymbol{\lambda}}$ on both sides of Eq (1). In matrix notation, ridge regression comes from perturbing the eigenvalues of $\mathbf{X}'\mathbf{X}$ as $d_i \rightarrow d_i + \boldsymbol{\lambda}$ while surrogate regression comes from perturbing the singular values of $\mathbf{X}$ as $\xi_i \rightarrow \sqrt{\xi_i^2 + \lambda}$. The raise estimators are based on perturbing a column $\mathrm{x}_j \rightarrow \tilde{\mathrm{x}}_j = \mathrm{x}_j + k_j \mathrm{e}_j$ where $\mathbf{e}_j$ is orthogonal to the span of the remaining resting columns.

The ridge and surrogate procedures do not require $\mathbf{X}$ to be of full rank. For example, with the surrogate transformation $\xi_i \rightarrow \sqrt{\xi_i^2 + \lambda}$, any zero singular value will be mapped to $\sqrt{\lambda} > 0$ with $\mathbf{X}_{\boldsymbol{\lambda}}$ now full rank. On the other hand, the raise procedure does require the columns of $\mathbf{X}$ to be independent as the crucial step $\mathrm{x}_1 \rightarrow \tilde{\mathrm{x}}_1(k_1) = \mathrm{x}_1 + k_1 \mathrm{e}_1$ moves $\mathbf{x}_1$ in the direction of the orthogonal complement of $Sp(\mathbf{X}_{[1]}) \subset Sp(\mathbf{X}) \subset \mathbb{R}^n$ in $Sp(\mathbf{X})$, the span of the other columns. Thus if $\mathrm{x}_1 \in Sp(\mathbf{X}_{[1]}) = Sp(\mathbf{X})$, then $Sp(\mathbf{X}_{[1]})^{\perp} \cap Sp(\mathbf{X}) = \{0\}$ and $\mathbf{x}_1$ cannot be raised. We now show how to evaluate each of these three estimators using the following example. Consider the following design matrix $\mathbf{X}$ with associated $\mathbf{Y}$

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 3 \\ 1 & 5 & 3 & 10 \\ 1 & 4 & 6 & 7 \\ 1 & 7 & 8 & 15 \\ 1 & 9 & 11 & 17 \end{bmatrix} \quad \mathbf{Y} = \begin{bmatrix} 4 \\ 19 \\ 16 \\ 34 \\ 40 \end{bmatrix}.$$

It is common practice (Belsley, 1986) to center the explanatory variables to improve the collinearity, with $\widehat{\beta}_0$ recovered from the relationship $\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\overline{x}_1 - \widehat{\beta}_2\overline{x}_2$. If we center, then the design matrix $\mathbf{CX}$ with associated $\mathbf{CY}$ become

$$\mathrm{CX} = \begin{bmatrix} -4.2 & -4.8 & -7.4 \\ -0.2 & -2.8 & -0.4 \\ -1.2 & 0.2 & -3.4 \\ 1.8 & 2.2 & 4.6 \\ 3.8 & 5.2 & 6.6 \end{bmatrix} \quad \mathrm{CY} = \begin{bmatrix} -18.6 \\ -3.6 \\ -6.6 \\ 11.4 \\ 17.4 \end{bmatrix}$$

for which the least squares estimators and variance inflation factors are respectively

$$\beta_{\mathbf{L}} = \begin{bmatrix} -3.063 \\ -0.773 \\ 1.018 \\ 2.286 \end{bmatrix} \quad VIF = \begin{bmatrix} 62.79 \\ 7.38 \\ 45.07 \end{bmatrix}.$$

The high variance inflation factor vector [62.79, 7.38, 45.07] is indicative of the high level of collinearity between the variables. A suggested acceptable level for *VIF* is 10 and we will show how to achieve this level for the ridge and the surrogate estimators. The condition numbers of $\mathbf{X}'\mathbf{X}$ and $(\mathbf{CX})'\mathbf{CX}$ respectively are 2296.15 and 454.39, demonstrating that centering has improved the collinearity.

Ridge Estimators

For our linear model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ view $\mathbf{X} = [\mathbf{X}_{[p]}, \mathbf{x}_p]$ with $\mathbf{x}_p$ the p^{th} column of $\mathbf{X}$ and $\mathbf{X}_{[p]}$ the matrix formed by the remaining columns. The variance inflation factors measure the effect of adding column $\mathbf{x}_p$ to $\mathbf{X}_{[p]}$. For notational convenience, we demonstrate with the last column p. An ideal column would be orthogonal to the previous columns with the entries in the off diagonal elements of the p^{th} row and p^{th} column of $\mathbf{X}'\mathbf{X}$ all zeros. Denote by $\mathbf{M}_p$ the idealized moment matrix

$$\mathbf{M}_p = \begin{bmatrix} \mathbf{X}'_{[p]}\mathbf{X}_{[p]} & \mathbf{0}_{p-1} \\ \mathbf{0}'_{p-1} & \mathbf{x}'_p\mathbf{x}_p \end{bmatrix}.$$

O'Driscoll and Ramirez (2015) show that

$$VIF(\widehat{\beta}_p) = \frac{\det(\mathbf{M}_p)}{\det(\mathbf{X}'\mathbf{X})} \tag{10}$$

In our example we set

$$\mathbf{Z}_\lambda'\mathbf{Z}_\lambda = (\mathbf{CX})'\mathbf{CX} + \lambda\mathbf{I_2} = \begin{bmatrix} 36.8+\lambda & 44.2 & 68.6 \\ 44.2 & 62.8+\lambda & 80.4 \\ 68.6 & 80.4 & 131.2+\lambda \end{bmatrix}$$

and to reduce the maximum variance inflation factor of 62.79 to 10 we solve

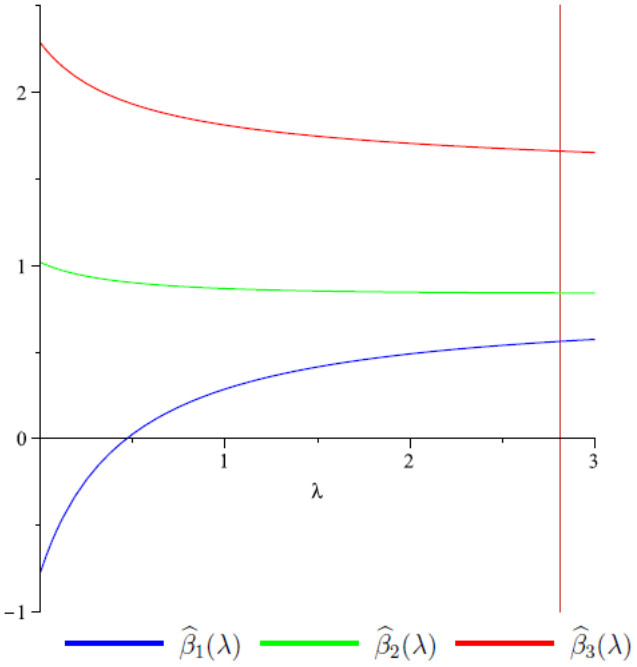
$$\frac{\det(\mathbf{M}_1)}{\det(\mathbf{Z}_\lambda'\mathbf{Z}_\lambda)} = 10 \tag{11}$$

for which $\boldsymbol{\lambda} = 2.810$ and the ridge estimators are [0.562, 0.842, 1.661].

A "ridge trace" plot of the behaviour of $\widehat{\beta}_1(\lambda), \widehat{\beta}_2(\lambda)$ and $\widehat{\beta}_3(\lambda)$ versus $\boldsymbol{\lambda} \geq 0$ is shown in Figure 4. McDonald (2010) states that "generally an analyst tends to a $\boldsymbol{\lambda}$ value where the trace has stabilized and major changes on the trace are to

the left of the chosen value for **λ**. For our choice of **λ** = 2.810 these characteristics are satisfied.

Figure 4. *Ridge Plots* $\widehat{\beta}_1(\lambda), \widehat{\beta}_2(\lambda), \widehat{\beta}_3(\lambda)$



Surrogate Estimators

By the Singular Value Decomposition Theorem there exist orthonormal matrices **U** and **V** such that **CX** = **UDV**′, where **D** is a diagonal matrix consisting of the singular values 14.853, 3.117 and 0.697 of **CX**. Then

$$\mathrm{D}_\lambda = \begin{bmatrix} \sqrt{220.60+\lambda} & 0 & 0 \\ 0 & \sqrt{9.714+\lambda} & 0 \\ 0 & 0 & \sqrt{0.485+\lambda} \end{bmatrix}$$

and to find the surrogate estimators we solve

$$\begin{aligned} \mathrm{VD}_\lambda \mathrm{U}'\mathrm{UD}_\lambda \mathrm{V}'\widehat{\beta_{\mathrm{S}}} &= \mathrm{VD}_\lambda \mathrm{U}'\mathrm{y} \\ \mathrm{VD}_\lambda^2 \mathrm{V}'\widehat{\beta_{\mathrm{S}}} &= \mathrm{VD}_\lambda \mathrm{U}'\mathrm{y}. \end{aligned} \qquad (12)$$

As in the case of the ridge estimator, to reduce the maximum variance inflation factor of 62.79 to 10, **λ** = 2.810 and the surrogate estimators are [0.195, 0.884, 1.843] with design matrix

$$\mathbf{CX_S} = \mathbf{VD}_\lambda\mathbf{U}' = \begin{bmatrix} -4.595 & -4.764 & -7.295 \\ 0.363 & -3.140 & -0.490 \\ -0.930 & 0.378 & -3.685 \\ 1.136 & 2.249 & 4.965 \\ 4.026 & 5.278 & 6.505 \end{bmatrix}.$$

We summarise our results in Table 1.

Table 1. *OLS, Ridge, and Surrogate Regression, Computed Parameters* $\boldsymbol{\lambda}$*, Estimated Coefficients* $\widehat{\beta}$*, Squared Lengths* $\widehat{\beta}'\widehat{\beta}$*,Condition Numbers* κ*, Maximum Variance Inflation Factor maxVIF,Mean Absolute Deviation for* $\mathbf{CX} - \mathbf{CX}_S$ *for Surrogate Design*

	OLS	Ridge	Surrogate
λ	0	2.810	2.810
$\widehat{\beta}$	$\begin{bmatrix} -0.773 \\ 1.019 \\ 2.286 \end{bmatrix}$	$\begin{bmatrix} 0.562 \\ 0.842 \\ 1.661 \end{bmatrix}$	$\begin{bmatrix} 0.195 \\ 0.885 \\ 1.844 \end{bmatrix}$
$\widehat{\beta}'\widehat{\beta}$	6.860	3.783	4.220
$\kappa(\mathbf{Z}_\lambda'\mathbf{Z}_\lambda)$	454.385	67.784	67.784
$maxVIF$	62.79	10.00	10.00
MAD	0	NA	0.24933

Raise Estimators

For the $n \times p$ matrix $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_p]$, the column span is denoted by $Sp(\mathbf{A})$, with $\mathbf{A}(j)$ denoting the j^{th} column vector $\mathbf{a}_j$ and $\mathbf{A}[j]$ denoting the $n\times (p-1)$ matrix formed by deleting $\mathbf{A}(j)$ from $\mathbf{A}$.

For the linear model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, central to a study of collinearity is the relationship between $\mathbf{X}(j)$ and $Sp(\mathbf{X}[j])$. We assume that the columns of $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_p)$ are centered.The raise estimators are based on perturbing a column $\mathrm{x}_j \rightarrow \widetilde{\mathrm{x}}_j = \mathrm{x}_j + k_j\mathrm{e}_j$ by a k_j multiple of a vector $\mathbf{e}_j$, which is orthogonal to the span of the remaining resting columns. We follow the notation from Garcia and Ramirez (201x). The regression of $\mathbf{x}_j$, viewed as the response vector using the remaining resting columns as the explanatory vectors, has an error vector $\mathbf{e}_j$ with the required properties. In this case, the projection matrix is

$$\mathbf{P_j} = \mathbf{X_{[j]}}(\mathbf{X_{[j]}}'\mathbf{X_{[j]}})^{-1}\mathbf{X_{[j]}}'$$

with error vector

$$\mathrm{e}_j = \mathrm{x}_j - \mathbf{P_j}\mathrm{x_j}.$$

The parameters vector $\mathbf{k} = (k_1, ..., k_p)'$ is to be chosen by the user. We illustrate how the raise estimators are constructed sequentially for the matrix $\mathbf{CX}$. Setting it follows that

$$\mathbf{P_1} = \begin{bmatrix} .423 & .073 & .147 & -.247 & -.395 \\ .073 & .484 & -.421 & .103 & -.238 \\ .147 & -.421 & .474 & -.224 & .024 \\ -.247 & .103 & -.224 & .190 & .179 \\ -.395 & -.238 & .024 & .179 & .431 \end{bmatrix} \quad \mathbf{e_1} = \begin{bmatrix} -0.288 \\ 0.417 \\ 0.213 \\ -0.506 \\ 0.164 \end{bmatrix}.$$

The raise estimators allow the user to specify, for each of the variables, a precision π_j that the data will retain during the raising stages by restricting the mean absolute deviation *MAD* inthe j^{th} column of $\mathbf{CX} - \widetilde{\mathbf{CX}}$ from

$$k_j \frac{1}{n} \sum_{i=1}^{n} |\mathbf{e}_{j,i}| = \pi_j. \tag{13}$$

Thus, given a specified precision $\pi_j > 0$, the user can raise column j in $\mathbf{CX}_{<1,\ldots,j>}$ to $\widetilde{\mathbf{cx}}_j(k_j) = \mathbf{cx}_j + \boldsymbol{\lambda}_j \mathbf{e}_j$, where k_j is solved from Eq. (13). From Table 1, we see that the mean absolute deviation *MAD* for $\mathbf{CX} - \widetilde{\mathbf{CX}}$ from the surrogate system is 0.24933. To compute a comparable raise system of estimators, we will set the precision $\pi_j = 0.24933$ in Eq. (13) and we firstly solve for $k_1 = 0.7843$ to raise $\mathbf{cx}_1 \rightarrow \widetilde{\mathbf{cx}}_1$.

The first raised design matrix is

$$\mathbf{CX}_{<1>} = \begin{bmatrix} -4.426 & -4.8 & -7.4 \\ 0.127 & -2.8 & -0.4 \\ -1.033 & 0.2 & -3.4 \\ 1.403 & 2.2 & 4.6 \\ 3.930 & 5.2 & 6.6 \end{bmatrix}$$

We now raise the vector cx2 from the regression of cx2 using the resting vectors from $\mathbf{CX}_{<1>}$.

$$\mathbf{cx_2} = \begin{bmatrix} -4.8 \\ -2.8 \\ 0.2 \\ 2.2 \\ 5.2 \end{bmatrix} \quad \mathbf{CX}_{[2]} = \begin{bmatrix} -4.426 & -7.4 \\ 0.127 & -0.4 \\ -1.033 & -3.4 \\ -1.403 & 4.6 \\ 3.930 & 6.6 \end{bmatrix}$$

$$\mathbf{P_2} = \begin{bmatrix} .558 & -.062 & .004 & -.007 & -.493 \\ -.062 & .052 & .124 & -.167 & .053 \\ .004 & .124 & .339 & -.457 & -.010 \\ -.007 & -.167 & -.457 & .616 & .015 \\ -.493 & .053 & -.010 & .015 & .435 \end{bmatrix} \quad \mathbf{e_2} = \begin{bmatrix} 0.279 \\ -2.884 \\ 1.556 \\ 0.361 \\ 0.688 \end{bmatrix}$$

From Eq.(13) $k_2 = 0.2161$ and the second raised matrix is

$$\mathbf{CX}_{<1,2>} = \begin{bmatrix} -4.426 & -4.740 & -7.4 \\ 0.127 & -3.423 & -0.4 \\ -1.033 & 0.536 & -3.4 \\ -1.403 & 2.278 & 4.6 \\ 3.929 & 5.349 & 6.6 \end{bmatrix}.$$

We finally raise the vector cx3 from the regression of $\mathbf{cx}_3$ using the resting vectors from $\mathbf{CX}_{<1,2>}$.

$$\mathbf{cx_3} = \begin{bmatrix} -7.4 \\ -0.4 \\ -3.4 \\ 4.6 \\ 6.6 \end{bmatrix} \quad \mathbf{CX}_{[3]} = \begin{bmatrix} -4.426 & -4.740 \\ 0.127 & -3.423 \\ -1.033 & 0.536 \\ -1.403 & 2.278 \\ 3.930 & 5.349 \end{bmatrix}$$

$$\mathbf{P_3} = \begin{bmatrix} .524 & -.099 & .161 & -.148 & -.438 \\ -.099 & .752 & -.369 & -.132 & -.153 \\ .161 & -.369 & .206 & .028 & -.026 \\ -.147 & -.132 & .028 & .077 & .175 \\ -.438 & -.153 & -.026 & .175 & .442 \end{bmatrix} \quad \mathbf{e_2} = \begin{bmatrix} 0.556 \\ -0.466 \\ -1.618 \\ 2.043 \\ -0.514 \end{bmatrix}$$

From Eq.(13) k_3 = 0.2398 and the final raised matrix is

$$\widetilde{\mathbf{CX}} = \mathbf{CX}_{<1,2,3>} = \begin{bmatrix} -4.426 & -4.740 & -7.267 \\ 0.127 & -3.423 & -0.512 \\ -1.033 & 0.536 & -3.788 \\ -1.403 & 2.278 & 5.090 \\ 3.930 & 5.349 & 6.477 \end{bmatrix}.$$

The *OLS* values from Table 1 are shown in Column 1 of Table 2 for comparisons.

Table 2. *OLS and Raise Regression with Precision* π_j *= 0.24933Computed Parameters* k_j*,Estimated Coefficients* $\widehat{\beta}$*, Squared Lengths* $\widehat{\beta}'\widehat{\beta}$*, Condition Numbers* κ*, Maximum Variance Inflation Factor maxVIF,*

	OLS	Step 1	Step 2	Step 3
π_j	0	0.24933	0.24933	0.24933
k_j	0	0.7843	0.2161	0.2399
$\widehat{\beta}$	$\begin{bmatrix} -0.773 \\ 1.018 \\ 2.286 \end{bmatrix}$	$\begin{bmatrix} -0.434 \\ 0.964 \\ 2.142 \end{bmatrix}$	$\begin{bmatrix} -0.266 \\ 0.793 \\ 2.160 \end{bmatrix}$	$\begin{bmatrix} 0.465 \\ 0.812 \\ 1.741 \end{bmatrix}$
$\widehat{\beta}'\widehat{\beta}$	6.860	5.704	5.361	3.908
$\kappa(\widetilde{\mathbf{CX}}'\widetilde{\mathbf{CX}})$	454.38	143.16	138.38	98.97
$maxVIF$	62.79	20.41	19.37	11.60

Conclusion

We have followed the standard practice of centering the explanatory variables which removes what is often dubbed as the nonessential collinearity. The ridge procedure perturbs the moment matrix $X'X \rightarrow X'X + \lambda \mathbf{I}_p$ but does not allow the user to compute the new design $\mathbf{X}$ and so the changes to $\mathbf{X}$ are unknown to the user. The surrogate procedure has the advantage of allowing the user to explicitly compute the new surrogate design $\mathbf{X}_S$ in terms of the singular values of $\mathbf{X}$. Thus the Mean Absolute Deviation $\|\mathbf{X} - \mathbf{X}_S\|_{MAD} = \frac{1}{np}\sum_{i,j}|\mathbf{X}_{i,j} - \mathbf{X}_{S\ ,i,j}|$ can be computed which gives the average change in the design $\mathbf{X} \rightarrow \mathbf{X}_S$. However, the surrogate procedure does not allow for the perturbations to differ for each explanatory variables. Since the ridge procedure cannot determine the new design, it seeks stability of $(\mathbf{X}'\mathbf{X})^{-1}$ only on the left-hand side of the normal equations $(\mathbf{X}'\mathbf{X})^{-1}\beta = \mathbf{X}'\mathbf{y}$, transforming the normal equations into the ridge equations $(\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}_p)^{-1}\beta = \mathbf{X}'\mathbf{y}$. Both the surrogate $\mathbf{X} \rightarrow \mathbf{X}_S$ and raised $\mathrm{X} \rightarrow \widetilde{\mathrm{X}}$ procedures use the modified matrix on both sides of the normal equations in the spirit of *OLS*.

It is important to note that the raise procedure does satisfy both objectives. It yields the explicit new design $\mathrm{X} \rightarrow \widetilde{\mathrm{X}}$ and thus the mean absolute deviations given in Eq.(13) are permitted to vary for each explanatory variable. Hence the user can set the mean absolute deviation to be small for variables which are known to be accurate and allow larger deviations for variables which are known to be less accurate. We believe this latter feature will be very useful with real data sets.

To compare the three ridge-type estimators, we first computed the *OLS* and ridge estimators. The ridge parameter was solved by reducing *maxVIF* down to 10.0, a common value used with variance inflation factors. Next the surrogate estimators were computed also by solving for the surrogate parameter that reduces *maxVIF* down to 10.0. The surrogate estimators are constructed so that both the ridge and surrogate moment matrices are identical with these choices of parameters. Since the surrogate design $\mathbf{X}_S$ can be explicitly computed, we are able to compute the Mean Absolute Deviation which measures the change to the original *OLS* design. Knowing the MAD allows for a fair comparison of the surrogate and raise estimators. The raised estimator $\beta(\lambda) = (\widetilde{\mathrm{X}}'\widetilde{\mathrm{X}})^{-1}\widetilde{\mathrm{X}}'\mathrm{y}$ is obtained from successive raising of the columns $\mathbf{X}$ using an elegant idea of perturbing a column of X in a direction orthogonal to the span of the other columns. This retains the *OLS* coefficient of determination R2. The raised parameter is solved by setting *MAD* to concur with the value from the surrogate procedure.

In our example the three methods gave similar results, where the main advantage of the surrogate over the ridge procedure is to allow the user to compute the new design. The main advantage of the raised over the surrogate procedure is that the latter allows the user to visualize the perturbations of the underlying model and easily control the amount of perturbations to the original design, retaining a specified precision in the data for each explanatory variable.

References

Belsley, D.A. (1986). Centering, the constant, first–differencing, and assessing conditioning. In *E. Kuh and D.A. Belsley, (Eds.),Model Reliability*. Cambridge: MIT Press, 117-153.

Davidov, O. (2006). Constrained Estimation and the Theorem of Kuhn-Tucker. In *Journal of Applied Mathematics and Decision Sciences*, Article ID 92970.

Garcia, C.B., Garcia, J. and Soto, J. (2011). The raise method: An alternative procedure to estimate the parameters in presence of collinearity, *Quality and Quantity*, 45, 403-423.

Garcia, J. and Ramirez, D. (201x). The successive raising estimator and its relation with the ridge estimator, in preparation.

Hoerl, A.E. (1959). Optimal solutions of many equations. *Chemical Engineering Progress, Symposium Series*, 55, 69-78.

Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12, 1 55-67.

Jensen, D.R. and Ramirez, D.E. (2008). Anomalies in the foundations of ridge regression. *Int. Stat. Rev*., 76, 89-105.

Jensen, D.R. and Ramirez, D.E. (2010). Tracking MSE efficiencies in ridge regression. *Advances and Applications in Statistical Sciences*, 1, 381-398.

McDonald, G.C. (2010). Tracing ridge regression coefficients, *Wiley Interdisciplinary Review: Computational Statistics*, 2, 695-793.

O'Driscoll, D. and Ramirez, D. (2015). Response surface design using the generalized variance inflation factors. *Cogent Mathematics*, 2, 1-11.

Tikhonov, A.N. (1943). On the stability of inverse problems. *Doklady Akademii Nauk SSSR*, 39 (5), 195-198.