O’Driscoll & Ramirez, *Cogent Mathematics* (2016), 3: 1144697
http://dx.doi.org/10.1080/23311835.2016.1144697

cogent mathematics

STATISTICS | RESEARCH ARTICLE

Mitigating collinearity in linear regression models using ridge, surrogate and raised estimators

Diarmuid O’Driscoll[1] and Donald E. Ramirez[2*]

Received: 21 October 2015
Accepted: 15 January 2016
First Published: 22 January 2016

*Corresponding author:
Donald E. Ramirez, Department of Mathematics, University of Virginia, Charlottesville, VA, USA
E-mail: der@virginia.edu

Reviewing editor:
Quanxi Shao, CSIRO, Australia

Additional information is available at the end of the article

Abstract: Collinearity in the design matrix is a frequent problem in linear regression models, for example, with economic or medical data. Previous standard procedures to mitigate the effects of collinearity included ridge regression and surrogate regression. Ridge regression perturbs the moment matrix $\mathbf{X}'\mathbf{X} \to \mathbf{X}'\mathbf{X} + k\mathbf{I}_p$, while surrogate regression perturbs the design matrix $\mathbf{X} \to \mathbf{X}_S$. More recently, the raise estimators have been introduced, which allow the user to track geometrically the perturbation in the data with $\mathbf{X} \to \widetilde{\mathbf{X}}$. The raise estimators are used to reduce collinearity in linear regression models by raising a column in the experimental data matrix, which may be nearly linear with the other columns, while keeping the basic *OLS* regression model. We give a brief overview of these three ridge-type estimators and discuss practical ways of choosing the required perturbation parameters for each procedure.

Subjects: Mathematical Statistics; Mathematics & Statistics; Science; Statistical Computing; Statistics; Statistics & Probability

Keywords: collinearity; ridge estimators; surrogate estimators; raise estimators

AMS Subject Classifications: 62J05; 62J07

1. Introduction

The standard linear regression model can be written as $\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$ with uncorrelated, zero-mean and homoscedastic errors $\boldsymbol{\varepsilon}$. Here $\mathbf{X}$ is a full rank $n \times p$ matrix containing the explanatory variables

ABOUT THE AUTHORS

Diarmuid O’Driscoll is the head of the Mathematics and Computer Studies Department at Mary Immaculate College, Limerick. He was awarded a Travelling Studentship for his MSc at University College Cork in 1977. He has taught at University College Cork, Cork Institute of Technology, University of Virginia, and Frostburg State University. His research interests are in mathematical education, errors in variables regression, ridge regression and design criteria. In 2014, he was awarded a Teaching Heroes Award by the National Forum for the Enhancement of Teaching and Learning (Ireland).

Donald E. Ramirez is a full professor in the Department of Mathematics at the University of Virginia in Charlottesville, Virginia. He received his PhD in Mathematics from Tulane University in New Orleans, Louisiana. His research is in harmonic analysis and mathematical statistics. His current research interests are in statistical outliers and ridge regression.

PUBLIC INTEREST STATEMENT

Collinearity is a frequent problem in statistical analysis of data, for example, with ordinary least square linear regression models of economic or medical data. Standard procedures to mitigate the effects of collinearity include ridge regression and surrogate regression. Ridge regression is based on a standard numerical technique that is used in computing an inverse of a nearly singular matrix. Surrogate regression is based on perturbing the data in a way to allow for more accurate numerical solutions. More recently, the raise estimators have been introduced. This technique also perturbs the data while allowing the researcher to track the changes in the data while retaining the basic ordinary least square regression model. We give a brief overview of these three ridge-type estimators and discuss practical ways of choosing the required perturbation parameters for each procedure. Our case study indicates an advantage for using the raise estimators.

and the response vector $\mathbf{y}$ is $n \times 1$ consisting of the observed data. The Ordinary Least Squared *OLS* estimators $\widehat{\boldsymbol{\beta}}_L$ are solutions of

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

given by

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

The solutions $\widehat{\boldsymbol{\beta}}_L$ are unbiased with variance matrix $V(\widehat{\boldsymbol{\beta}}_L) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. For convenience, we take $\sigma^2 = 1$. The *OLS* solutions require that $(\mathbf{X}'\mathbf{X})^{-1}$ be accurately computed.

2. Ridge and surrogate estimators

With economic or medical data, the predictor variables in the columns of $\mathbf{X}$ may have a high level of collinearity; that is, there may be a nearly linear relationship among the predictor variables. In this case, $\mathbf{X}'\mathbf{X}$ in Equation (1) is nearly singular and thus $(\mathbf{X}'\mathbf{X})^{-1}$ will be numerically difficult to evaluate. It was observed by Riley (1955) that the perturbed matrix $\mathbf{X}'\mathbf{X} + k\mathbf{I}_p$ with $k > 0$ is better conditioned than the matrix $\mathbf{X}'\mathbf{X}$ and he suggested using the perturbed matrix in Equation (1). With $k > 0$ large enough, $(\mathbf{X}'\mathbf{X} + k\mathbf{I}_\mathbf{p})^{-1}$ can be accurately computed with standard numerical procedures. Using $\mathbf{X}'\mathbf{X} \to \mathbf{X}'\mathbf{X} + k\mathbf{I}_p$, Hoerl (1964) dubbed this procedure *ridge regression* with *ridge estimators*

$$\widehat{\boldsymbol{\beta}}_R(k) = (\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{y}. \tag{3}$$

Near dependency among the columns of $\mathbf{X}$ causes ill-conditioning in $\mathbf{X}'\mathbf{X}$ which results in *OLS* solutions with inflated squared lengths $||\widehat{\boldsymbol{\beta}}_L||^2$, with $\widehat{\boldsymbol{\beta}}_L$ of questionable signs (±) and with $\widehat{\boldsymbol{\beta}}_L$ being "very sensitive to small changes in $\mathbf{X}$" (Belsley, 1986). With ill-conditioning in $\mathbf{X}'\mathbf{X}$, the *OLS* solutions at $k = 0$ in Equation (3) are known to be unstable with a slight movement away from $k = 0$ giving completely different estimates of the coefficients $\boldsymbol{\beta}$.

In *The International Encyclopedia of Statistical Science,* Hadi (2011) discusses two standard remedies for addressing collinearity in linear regression; namely (1) the *ridge system* $\{(\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)\beta = \mathbf{X}'\mathbf{y};\, k \geq 0\}$ (Hoerl & Kennard, 1970) with solutions $\{\widehat{\boldsymbol{\beta}}_R(k);\, k \geq 0\}$ and (2) the *surrogate system* $\{(\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)\beta = (\mathbf{X}_k'\mathbf{X}_k)\boldsymbol{\beta} = \mathbf{X}_k'\mathbf{y};\, k \geq 0\}$ (Jensen & Ramirez, 2008) with solutions $\{\widehat{\boldsymbol{\beta}}_S(k); k \geq 0\}$. The ridge estimators come from modifying $\mathbf{X}'\mathbf{X} \to \mathbf{X}'\mathbf{X} + k\mathbf{I}_p$ on the left side of Equation (1) while the Jensen and Ramirez surrogate estimators modify the design matrix $\mathbf{X} \to \mathbf{X}_k$ on both sides of Equation (1). In matrix notation, ridge regression comes from perturbing the eigenvalues of $\mathbf{X}'\mathbf{X}$ as $\lambda_i \to \lambda_i + k$, while surrogate regression comes from perturbing the singular values of $\mathbf{X}$ as $\xi_i \to \sqrt{\xi_i^2 + k}$. From the singular value decomposition $\mathbf{X} = PD(\xi_i)\mathbf{Q}'$, the surrogate design is $\mathbf{X}_k = \mathbf{P}D(\sqrt{\xi_i^2 + k})\mathbf{Q}'$, with $\mathbf{D}$ a diagonal matrix of dimension $n \times p$, the columns of $\mathbf{P}$ the left-singular vectors and the columns of $\mathbf{Q}$ the right singular vectors. The surrogate transformation $\mathbf{X} \to \mathbf{X}_k$ preserves the ridge moments, with $\mathbf{X}_k'\mathbf{X}_k = \mathbf{X}'\mathbf{X} + k\mathbf{I}_p$ allowing for comparison between the two methods. Ridge regression has a long history of use in the statistical literature. The earliest detailed expositions of ridge estimators are found in Marquardt (1963) and Hoerl and Kennard (1970), with Marquardt (1963) acknowledging that Levenberg (1944) had observed that a perturbation of the diagonal improved convergence in steepest descent algorithms. The history of the early use of matrix diagonal increments in statistical problems is given in the article by Piegorsch and Casella (1989).

To alleviate the problems inherent with a singular value, say ξ_p, which is indicating collinearity in $\mathbf{X}$, the surrogate transformation converts $\xi_p \to \sqrt{\xi_p^2 + k}$ moving the singular value away from zero. Principal Component Regression (PCR) does the opposite and replaces ξ_p with 0 and regresses $\mathbf{Y} = \mathbf{P}D(\xi_1, \ldots, \xi_{p-1}, 0)\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\beta} = Q\alpha$. Hadi and Ling (1998) have noted "that it is possible for the PCR to fail miserably." Their example is constructed with the response variable $\mathbf{Y}$ being highly correlated with the deleted eigenvector associated with the deleted singular value. This deletion results in the remaining explanatory variables being unable to provide a good fit for the response variable.

Since ridge regression is based on a *numerical analysis* technique, the ridge estimators may lack desirable *statistical* properties. Three such desirable statistical properties follow.

(1) The *condition number* for a square $p \times p$ matrix $\mathbf{A}$ is a measure of the ill-conditioning in $\mathbf{A}$ and is defined as the ratio of the largest to smallest eigenvalues, denoted $\kappa(\mathbf{A}) = \lambda_1/\lambda_p$. Since perturbation procedures are designed to *improve* the regression model, one would expect that as $k \to \infty$ that $\kappa(V(\widehat{\boldsymbol{\beta}}_R(k)) \to 1$. However, as shown in Jensen and Ramirez (2010a), $\kappa(V(\widehat{\boldsymbol{\beta}}_R(k)) \to \kappa(V(\widehat{\boldsymbol{\beta}}_R(0))$. Initially, as k increases, the ill-conditioning in the variance matrix starts to get better but then returns to the original (bad) value. However, the surrogate system does have the desirable monotone property that $\kappa(V(\widehat{\boldsymbol{\beta}}_S(k)) \to 1$ as $k \to \infty$. This allows the user of surrogate estimators to be assured that, regardless of the chosen value for k, the variance matrix for the surrogate estimators will be more "orthogonal" than the original *OLS* variance matrix.

(2) Denote $V(\widehat{\boldsymbol{\beta}}_L) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2\mathbf{V}$ so v_{jj} is the actual variance for $\widehat{\boldsymbol{\beta}}_{L\,j}$ and denote $\mathbf{X}'\mathbf{X} = \mathbf{W}$. An "ideal" predictor variable in column j would be orthogonal to the other predictor variables in $\mathbf{X}$, with $\mathbf{W}$ being zero for all off-diagonal values in the j^{th} row and j^{th} column. In this "ideal" case, the "ideal" variance for $\widehat{\boldsymbol{\beta}}_{L\,j}$ would be $\sigma^2(\mathbf{W})^{-1}[j,j] = \sigma^2 w_{jj}^{-1}$. The *Variance Inflation Factors* (*VIFs*) of $\widehat{\beta}_L = [\widehat{\beta}_{L1}, \ldots, \widehat{\beta}_{Lp}]'$ are given by $\{VIF(\widehat{\beta}_{L\,j}) = v_{jj}/w_{jj}^{-1}; 1 \le j \le p\}$; i.e. the ratios of actual variances to "ideal" variances had the columns of $\mathbf{X}$ been orthogonal, with $VIF(\widehat{\beta}_{L\,j}) = 1$ for the ideal orthogonal case. Marquardt and Snee (1975) have identified *VIF* as "the best single measure of the conditioning of the data." Again since perturbation procedures are designed to *improve* the regression model, one would expect that as $k \to \infty$ that $VIF(V(\widehat{\boldsymbol{\beta}}_{R\,j}(k)) \to 1$. Jensen and Ramirez (2010a) also showed that $VIF(V(\widehat{\boldsymbol{\beta}}_{R\,j}(k)) \to VIF(V(\widehat{\boldsymbol{\beta}}_{R\,j}(0))$ for the ridge estimators but that $VIF(V(\widehat{\boldsymbol{\beta}}_{S\,j}(k)) \to 1$ as $k \to \infty$ for the surrogate estimators, resulting in less collinearity between the surrogate estimators than exists between the *OLS* estimators.

(3) Hoerl and Kennard (1970) established that the ridge estimators satisfy the *MSE Admissibility Condition* assuring an improvement in Mean Squared Error $MSE(\widehat{\boldsymbol{\beta}}_R(k))$ for some $k \in (0, \infty)$. With $\widehat{\mathbf{y}}_R(k)$, the predicted values for ridge regression, the statistic $MSE(\widehat{\boldsymbol{\beta}}_R(k)) = \sum_j (y_j - \hat{y}_{R\,i}(k))^2 = ||\mathbf{y} - \widehat{\mathbf{y}}_R(k)||^2$ measures how close the predicted values in the ridge regression model are to the observed values. However, Jensen and Ramirez (2010b) have shown the existence of cross-over values k_0 for which, if $k > k_0$ then $MSE(\widehat{\boldsymbol{\beta}}_R(k)) > MSE(\widehat{\boldsymbol{\beta}}_L(k))$, indicating that the ridge model should not be used. The Hoerl and Kennard (1970) result assures that for some positive value of k, the ridge model is an improved model. Jensen and Ramirez (2010a) have shown that for any $k \in (0, \infty)$ the corresponding result holds for surrogate estimators. A further improvement with surrogate estimators is given by $MSE(\widehat{\boldsymbol{\beta}}_S(k)) \le MSE(\widehat{\boldsymbol{\beta}}_R(k))$; that is, for any value of k, the surrogate estimators have predicted values closer to the original data than the ridge estimators. As the ridge and surrogate estimators are not equivariant under scaling, the common convention is to scale $\mathbf{X}'\mathbf{X}$ to *correlation form* with the explanatory variables centered and scaled to unit length.

Remark 1 Scaling $\mathbf{X}'\mathbf{X}$ to correlation form can lead to some anomalies. as noted in Jensen and Ramirez (2008). For example, the map $k \to ||\widehat{\boldsymbol{\beta}}_R(k)||^2$ is known to be monotonically decreasing with $\mathbf{X}$ centered but unscaled. Using Proc Reg in SAS with the Ridge option, this monotone property can be lost as the original $\mathbf{X}'\mathbf{X}$ moment matrix is (1) scaled into correlation form and (2) the ridge estimators are computed using the correlation form for $\mathbf{X}'\mathbf{X}$ and (3) the ridge solutions are mapped back into the original scale. This scaling-rescaling can cause $k \to ||\widehat{\boldsymbol{\beta}}_R(k)||^2$ to lose its monotonicity as in the example in Jensen and Ramirez (2008).

Remark 2 Let $\mathbf{X}$ be mean-centered. Let $\mathbf{D}^2$ be the diagonal matrix with entries $1/\mathbf{X}'\mathbf{X}_{j,j}$, $1 \le j \le p$, then the scaling $\mathbf{X} \to \mathbf{XD}$ has $(\mathbf{XD})'(\mathbf{XD})$ in correlation form., that is with diagonal entries all having value one. This is the scaling we have used. Sardy (2008) has suggested a covariance-based scaling using the diagonal matrix $\mathbf{D}_\Sigma^2$ with entries $(\mathbf{X}'\mathbf{X}\,)^{-1}_{j,j}$, $1 \le j \le p$. We note that in this case $(\mathbf{XD}_\Sigma)'(\mathbf{XD}_\Sigma)$ has diagonal entries which are the variance inflation factors $VIF(\hat{\beta}_j)$. The variance inflation factors are the ratios of the variances of $\hat{\beta}_j$ to the "ideal" variances of $\hat{\beta}_j$ assuming the explanatory variables are orthogonal; that is, $VIF(\hat{\beta}_j) = (\mathbf{X}'\mathbf{X})^{-1}_{j,j}/(1/\mathbf{X}'\mathbf{X}_{j,j}) = (\mathbf{X}'\mathbf{X})^{-1}_{j,j}\mathbf{X}'\mathbf{X}_{j,j} = (\mathbf{X}'\mathbf{X})^{-1/2}_{j,j}\mathbf{X}'\mathbf{X}_{j,j}(\mathbf{X}'\mathbf{X})^{-1/2}_{j,j}$. In our Case Study, $p = 2$ so $VIF(\hat{\beta}_1) = VIF(\hat{\beta}_2)$ and $c(\mathbf{XD})'(\mathbf{XD}) = (\mathbf{XD}_\Sigma)'(\mathbf{XD}_\Sigma)$ with c the common value of the variance inflation factors.

Remark 3 When the regression model retains the parameter β_0 for the constant term with the design matrix $\mathbf{X}$ containing a unit constant column, the user needs to be careful with defining $VIF(\hat{\beta}_j)$ when the data have not been mean-centered. In short, $VIF(\hat{\beta}_j)$ is based on comparing the (j, j) entry of the variance matrix to the corresponding entry of an "ideal" covariance matrix. The inverse of the "ideal" covariance matrix is denoted as the "ideal" moment matrix. The "ideal" $\hat{\beta}_j$ is uncorrelated with the other explanatory variables $\hat{\beta}_i$, $0 < i \ne j$. Thus, the constraints on the "ideal" covariance matrix are that (1) the off-diagonal (i, j) and (j, i) entries for $cov(\hat{\beta}_i, \hat{\beta}_j)$ are zero where $0 < i \ne j$. Note that the "ideal" covariance matrix is not a diagonal matrix as the entries relating to $\hat{\beta}_0$ in the first row and column are retained as the data have not been centered. Additionally, the constraints on the "ideal" moment matrix are that (2) the entries in first row and first column are the first order moments determined from the data and (3) the entries down the diagonal (j, j) with $j \ge 0$ are the second order moments determined from the data. Jensen and Ramirez (2013) have given an easy to compute algorithm for computing the "ideal" covariance matrix that satisfies constraints (1), (2) and (3).

The variance inflation factors, which are the standard measure for collinearity, have a geometric interpretation which allows them to be conveniently computed as a ratio of determinants. We assume that the variables are *centered*. Reorder $\mathbf{X} = [\mathbf{X}_{[p]}, \mathbf{X}_{(p)}]$ with $\mathbf{X}_{(p)} = \mathbf{x}_p$ the p^{th} column and $\mathbf{X}_{[p]}$, the design matrix $\mathbf{X}$ without the p^{th} column, dubbed the *resting columns*. Garcia, Garcia and Soto (2011) introduced the *metric number* to measure the effect of adding the last column $\mathbf{X}_{(p)}$ to the resting columns $\mathbf{X}_{[p]}$. An ideal p^{th} column would be orthogonal to the other 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, with idealized $\mathbf{X}'\mathbf{X}$ moment matrix

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

The metric number is defined by $MN(\mathbf{x}_p) = \left(\det(\mathbf{X}'\mathbf{X})/\det(\mathbf{M}_p)\right)^{1/2}$ and it measures the effect of enlarging the design matrix with the adding of the p^{th} exploratory column. The metric number is easy to compute and is functionally equivalent to the *VIF* statistics with

$$VIF(\hat{\beta}_p) = \frac{\det(\mathbf{M}_p)}{\det(\mathbf{X}'\mathbf{X})},$$

for example, O'Driscoll and Ramirez (2015).

In spite of the established usage of ridge regression, it is now known that the surrogate estimators have superior statistical properties over the ridge estimators. Indeed, for their statistical analysis, Woods et al. (2012) used the Jensen–Ramirez surrogate estimates for modelling of diabetes in stock rats.

A crucial question for both the ridge estimators and the surrogate estimators is: What value of k should be used? McDonald (2009, 2010) has suggested that k can be determined by controlling the correlation between the observed values and the predicted values from ridge regression. We extend this methodology to surrogate regression and will compare the two procedures.

McDonald (2009, 2010) showed that the square of the correlation coefficient $R^2(\hat{\boldsymbol\beta}_R(k))$ between the observed values $\mathbf{y}$ and the ridge predicted values $\hat{\mathbf{y}}_R(k) = \mathbf{X}\hat{\boldsymbol\beta}_R(k)$ is a monotone decreasing function in the ridge parameter k. The corresponding result for the square of the correlation coefficient $R^2(\hat{\boldsymbol\beta}_S(k))$ of the observed values $\mathbf{y}$ and the surrogate predicted values $\hat{\mathbf{y}}_S(k) = \mathbf{X}\hat{\boldsymbol\beta}_S(k)$ for the surrogate regression is a monotone decreasing function in the surrogate parameter k, as shown in Garcia and Ramirez (in press). This allows the user to determine a unique value for k by controlling the decrease in correlation between the observed and predicted values. The user can set a lower bound for the reduction in $R^2(\hat{\boldsymbol\beta}_R(k))$ and $R^2(\hat{\boldsymbol\beta}_S(k))$ and numerically compute the associated ridge and surrogate parameters, For example, to preserve 95% of the *OLS* correlation, we solve $R^2(k) = 0.95R^2(0)$. With the computed value for k, we can measure the reduction in collinearity using the *VIF* statistic or using the condition number κ of $\mathbf{X}'\mathbf{X} + k\mathbf{I}_p$. For our case study, we use the example in McDonald (2010) which is known to have severe collinearity. We report the improvements in collinearity for both methods.

3. Raise estimators

We assume that the columns of $\mathbf{X} = \left(\mathbf{x}_1, \mathbf{x}_2, \dots, x_p\right)$ are centered and standardized, that is, $\mathbf{X}'\mathbf{X}$ is in correlation form with $||\mathbf{x}_j||^2 = 1$. For the $n \times p$ matrix $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_p]$, 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 $\mathbf{y} = \mathbf{X}\beta + \boldsymbol\varepsilon$, central to a study of collinearity is the relationship between $\mathbf{X}_{(j)}$ and $Sp(\mathbf{X}_{[j]})$.

The raise estimators are based on perturbing a column $\mathbf{x}_j \to \tilde{\mathbf{x}}_j = \mathbf{x}_j + \lambda_j\mathbf{e}_j$ by a λ_j multiple of a vector $\mathbf{e}_j$ orthogonal to the span of the remaining *resting* columns. We follow the notation from Garcia and Ramirez (in press). 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. The raise estimators are constructed sequentially as follows.

- Step 1: We raise the vector $\mathbf{x}_1$ from the regression of $\mathbf{x}_1$ using the resting vectors $\mathbf{X}_{[1]} = \left(\mathbf{x}_2, \mathbf{x}_3, \dots x_p\right)$. From this regression, we take the error vector $\mathbf{e}_1$ with $\mathbf{e}_1 \perp Sp(\mathbf{X}_{[1]})$ to construct $\tilde{\mathbf{x}}_1(\lambda_1) = \mathbf{x}_1 + \lambda_1\mathbf{e}_1$. The raised design matrix is denoted $\mathbf{X}_{<1>} = \left(\tilde{\mathbf{x}}_1(\lambda_1), \mathbf{x}_2, \dots, x_p\right)$.
- Step j: we raise the vector $\mathbf{x}_j$ from the regression of $\mathbf{x}_j$ using the resting vectors from $\mathbf{X}_{<1,\dots,j-1>}$, namely $\mathbf{X}_{<1,\dots,j-1>[j]} = \left(\tilde{\mathbf{x}}_1(\lambda_1), \dots, \tilde{x}_{j-1}(\lambda_{j-1}), \mathbf{x}_{j+1}, \dots, x_p\right)$. From this regression, we take the residual vector $\mathbf{e}_j$ with $\mathbf{e}_j \perp Sp(\mathbf{X}_{<1,\dots,j-1>[j]})$ to construct $\tilde{\mathbf{x}}_j(\lambda_j) = \mathbf{x}_j + \lambda_j\mathbf{e}_j$. The raised design matrix is denoted $\mathbf{X}_{<1,\dots,j>} = \left(\tilde{\mathbf{x}}_1(\lambda_1), \dots, \tilde{x}_j(\lambda_j), \mathbf{x}_{j+1}, \dots, x_p\right)$.
- Step p: we raise the vector $\mathbf{x}_p$ from the regression of $\mathbf{x}_p$ with the resting vectors from $\mathbf{X}_{<1,\dots,p-1>}$, namely $\mathbf{X}_{<1,\dots,p-1>[p]} = \left(\tilde{\mathbf{x}}_1(\lambda_1), \dots, \tilde{x}_{p-1}(\lambda_{p-1})\right)$. Then, we take the residual $\mathbf{e}_p$ with $\mathbf{e}_p \perp Sp(\mathbf{X}_{<1,\dots,p-1>[p]})$ to construct $\tilde{\mathbf{x}}_p = \mathbf{x}_p + \lambda_p\mathbf{e}_p$. The raised design matrix is denoted $\mathbf{X}_{<1,\dots,p>} = \left(\tilde{\mathbf{x}}_1(\lambda_1), \dots, \tilde{x}_p(\lambda_p)\right)$ with parameters vector $\boldsymbol\lambda = (\lambda_1, \dots, \lambda_p)'$ to be chosen by the user. For convenience, we denote the final raise design $\mathbf{X}_{<1,\dots,p>}$ by $\widetilde{\mathbf{X}}$.

There is a monotone relationship between the variance inflation factors, VIF_j, and the angle between $(\mathbf{x}_j,\ Sp(\mathbf{X}_{[j]}))$, for example, Jensen and Ramirez (2013, Theorem 4). Let $\mathbf{P}_{[j]} = \mathbf{X}_{[j]}(\mathbf{X}'_{[j]}\mathbf{X}_{[j]})^{-1}\mathbf{X}'_{[j]}$ be the projection operator onto the subspace $Sp(\mathbf{X}_{[j]}) \subset \mathbb{R}^n$ spanned by the columns of the reduced (or relaxed) matrix $\mathbf{X}_{[j]}$. From the geometry of the right triangle formed by $(\mathbf{x}_j, \mathbf{P}_{[j]}\mathbf{X}_{(j)})$, it can be shown that the angle θ_j between $\mathbf{x}_j$ and $\mathbf{P}_{[j]}\mathbf{X}_{(j)}$ satisfies $\cos(\theta_j) = ||\mathbf{P}_{[j]}\mathbf{x}_j||/||\mathbf{x}_j||$ and similarly the angle between $\widetilde{\mathbf{x}}_j$ and $\mathbf{P}_{[j]}\mathbf{X}_{(j)}$ satisfies $\cos(\widetilde{\theta}_j) = ||\mathbf{P}_{[j]}\mathbf{x}_j||/||\widetilde{\mathbf{x}}_j||$ since $\mathbf{e}_j \perp \mathbf{P}_{[j]}\mathbf{x}_j$. Each Variance Inflation Factor, VIF_j, for $\widetilde{\mathbf{X}}$ is functionally related to the angle $\widetilde{\theta}_j$ by the rule $\widetilde{\theta}_j = \arccos(\sqrt{1 - 1/VIF_j})$, for example Jensen and Ramirez (2013). Thus as $\lambda_j \to \infty$, $\widetilde{\theta}_j \to 90^\circ$ and the variance inflation factor VIF_j converges to one indicating that collinearity is being diminished, as in Garcia et al. (2011, Theorem 4.2).

Some desirable properties of the raised regression method are as follows.

(1) Raising a column vector in $\mathbf{X}$ does not effect the basic *OLS* regression model as the raised vector remains in the original $Sp(\mathbf{X})$, $\mathbf{e}_j = \mathbf{X}_{(j)} - \mathbf{P}_{[j]}\mathbf{X}_{(j)} \in Sp(\mathbf{X})$ so $Sp(\widetilde{\mathbf{X}}) = Sp(\mathbf{X})$, as shown in Garcia et al. (2011).

(2) Garcia et al. (2011) has shown that the raise estimators satisfy the *MSE Admissibility Condition* assuring an improvement in Mean Squared Error $MSE(\widetilde{\boldsymbol{\beta}}(\lambda))$ for some $\lambda \in (0, \infty)$ and thus the raise estimators can be said to be of ridge-type.

(3) The *VIFs* associated with the raise estimators are monotone functions, decreasing with λ_j, see Garcia, Garcia, López Martin, and Salmeron (2015).

(4) Starting with $\mathbf{X}'\mathbf{X}$ in correlation form with

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1p} \\ \rho_{12} & 1 & \rho_{23} & \cdots & \rho_{2p} \\ \rho_{13} & \rho_{23} & 1 & \cdots & \rho_{3p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho_{1p} & \rho_{2p} & \rho_{3p} & \cdots & 1 \end{pmatrix}$$

results in the final raising matrix $\widetilde{\mathbf{X}}$ having moment matrix

$$\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} = \begin{pmatrix} 1+\lambda_1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1p} \\ \rho_{12} & 1+\lambda_2 & \rho_{23} & \cdots & \rho_{2p} \\ \rho_{13} & \rho_{23} & 1+\lambda_3 & \cdots & \rho_{3p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho_{1p} & \rho_{2p} & \rho_{3p} & \cdots & 1+\lambda_p \end{pmatrix} \tag{4}$$

Thus, the raised regression perturbation matrix is equivalent to a *generalized ridge* regression perturbation matrix. And conversely, any generalized ridge regression matrix has a corresponding raised regression matrix as in Garcia and Ramirez (in press).

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* in the j^{th} column of $\mathbf{X} - \widetilde{\mathbf{X}}$ from

$$\lambda_j \frac{1}{n} \sum_{i=1}^{n} |\mathbf{e}_{j,i}| = \pi_j. \tag{5}$$

Thus, given a specified precision $\pi_j > 0$, the user can raise column j in $\mathbf{X}_{<1,\ldots,j>}$to $\widetilde{\mathbf{x}}_j(\lambda_j) = \mathbf{x}_j + \lambda_j \mathbf{e}_j$, where λ_j is solved from Equation (5). The precision values should be based on the researcher's belief in the accuracy of the data. The raised parameters λ_j are thus constrained to assure that the original data have not been perturbed more than what the researcher has permitted.

Table 1. *OLS*, ridge, and surrogate regression with squared correlation $R^2(\mathbf{y},\hat{\mathbf{y}}(k))$, computed parameters k, estimated coefficients $\hat{\boldsymbol{\beta}}$, squared lengths $\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}$, condition numbers κ, variance inflation factors *VIF*, and mean absolute deviation for $\mathbf{X}-\mathbf{X}_k$ for surrogate design

	OLS	Ridge	Surrogate
$R^2(\mathbf{y},\hat{\mathbf{y}}(k))$	0.3249	0.3086	0.3086
k	0	0.01822	0.05775
$\hat{\boldsymbol{\beta}}$	$\begin{bmatrix} 3.0255 \\ -3.1539 \end{bmatrix}$	$\begin{bmatrix} 1.3886 \\ -1.5158 \end{bmatrix}$	$\begin{bmatrix} 1.0860 \\ -1.212 \end{bmatrix}$
$\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}$	19.10	4.23	4.18
$\kappa(\mathbf{X}'\mathbf{X}+k\mathbf{I}_p)$	122.76	58.23	27.62
VIF	31.19	15.06	12.53
MAD	0	none	0.009158

Table 2. *OLS* and raise regression with precision $\pi_j = 0.009158$ squared correlation $R^2(\mathbf{y},\hat{\mathbf{y}}(k))$, computed parameters λ_j, estimated coefficients $\hat{\boldsymbol{\beta}}$, squared lengths $\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}$, condition numbers κ, variance inflation factors *VIF*, and mean absolute deviation for $\mathbf{X}-\widetilde{\mathbf{X}}$ for raise design

	OLS	Step 1	Step 2
π_j	0	0.009158	0.009158
$R^2(\mathbf{y},\hat{\mathbf{y}}(\lambda))$	0.3249	0.3169	0.3147
λ	0	0.5671	0.3898
$\hat{\boldsymbol{\beta}}$	$\begin{bmatrix} 3.0255 \\ -3.1539 \end{bmatrix}$	$\begin{bmatrix} 1.9307 \\ -2.0767 \end{bmatrix}$	$\begin{bmatrix} 1.3831 \\ -1.4942 \end{bmatrix}$
$\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}$	19.10	8.04	4.15
$\kappa(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})$	122.76	51.19	27.43
VIF	31.19	13.29	7.36
θ_j	10.31°	15.92°	21.62°
MAD	0	0.009158	0.009158

Remark 4 The ridge and surrogate procedures do not require $\mathbf{X}$ to be of full rank. For example, with the surrogate transformation $\xi_i \to \sqrt{\xi_i^2 + k}$ any zero singular value will be mapped to $\sqrt{k} > 0$ with $\mathbf{X}_k$ now full rank. On the other hand, the raise procedure does require the columns of $\mathbf{X}$ to be independent as the crucial step $\mathbf{x}_1 \to \tilde{\mathbf{x}}_1(\lambda_1) = \mathbf{x}_1 + \lambda_1\mathbf{e}_1$ 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 $\mathbf{x}_1 \in Sp(\mathbf{X}_{[1]}) = Sp(\mathbf{X})$ then $Sp(\mathbf{X}_{[1]})^\perp \cap Sp(\mathbf{X}) = \{0\}$ and $\mathbf{x}_1$ cannot be raised.

4. Case study

Our case study is the numerical example in McDonald (2010). Here, $n = 60$ and $p = 2$ with $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2]$ with $\mathbf{x}_1$ the nitrogen oxide pollution potential and $\mathbf{x}_2$ the hydrocarbon pollution potential and $\mathbf{y}$ the total mortality rate in 60 US metropolitan areas. The original data-set had 15 explanatory variables. Following McDonald (2010), we concentrate on the two variables which have the highest correlation $p = 0.9838$. Since, $\mathbf{X}$ is assumed to contain only explanatory variables, the vectors $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{y}$ are all mean-centered and scaled to have unit length.

Table 1 reports results for *OLS*, ridge and surrogate regression for the case study. The squared correlation between the original data $\mathbf{y}$ and the predicted values $\hat{\mathbf{y}}_L(k)$ for *OLS* is 0.3249 . The perturbation parameter for ridge and surrogate are chosen to retain 95% of this value for the squared correlation between $\mathbf{y}$ and $\widehat{\mathbf{y}}_R = \mathbf{X}\beta_R(k)$ for the ridge parameter and between $\mathbf{y}$ and $\widehat{\mathbf{y}}_S = \mathbf{X}\beta_S(k)$ for the surrogate parameter. Thus, both methods have the same small decrease in $R^2(\mathbf{y}, \widehat{\mathbf{y}}(k))$ down to 0.3086 shown in Row 1 of Table 1 and with the associated parameters in Row 2 of Table 1. This allows us to compare the improvement in collinearity between the two procedures. The estimated coefficients are shown in Row 3 of Table 1.

Ridge-type procedures are designed to (1) decrease the squared length of the estimated coefficient $\boldsymbol{\beta}'\boldsymbol{\beta}$ which is given in Row 4 of Table 1; (2) to decrease the condition number $\kappa(\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)$ of the matrix which needs to be inverted which is given in Row 5 of Table 1; (3) to decrease the variance inflation factors *VIF* given in Row 6. Since $p = 2$, both *VIFs* have a common value so only one value appears in Row 6. For each of these three criteria, surrogate regression is shown to be a superior procedure achieving a model with smaller collinearity with comparable loss of squared correlation $R^2(\mathbf{y}, \widehat{\mathbf{y}}(k))$.

The standard method for computing *VIF* for ridge regression in correlation form follows the procedure suggested by Marquardt (1970), which is to use the values on the main diagonal values of $(\mathbf{X}'\mathbf{X} + k\mathbf{I}_\mathbf{p})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X} + k\mathbf{I}_\mathbf{p})^{-1}$. Although this is the correct expression for $k = 0$, it has been shown to be in error for $k > 0$ by Garcia et al. (2015) as the Marquardt expression allows inadmissible values less than one. Thus, we have used Equation (22) and (25) in Garcia et al. (2015) to compute the corrected values for *VIF* for ridge regression.

From Table 1, we see that the mean absolute deviation *MAD* for $\mathbf{X} - \widetilde{\mathbf{X}}$ from the surrogate system is 0.009158. To compute a comparable raise system of estimators, we will set the precision $\pi_j = 0.009158$ in Equation (5). The *OLS* values from Table 1 are shown in Column 1 of Table 2 for comparisons. Using $\pi_1 = 0.009158$ in Step 1, we solve for $\lambda_1 = 0.5671$ to raise $\mathbf{x}_1 \rightarrow \widetilde{\mathbf{x}}_1$. With this value, the squared lengths $\widehat{\boldsymbol{\beta}}'\widehat{\boldsymbol{\beta}} = 8.04$, $\kappa = 51.19$ and $VIF = 13.29$ all showing an improvement in collinearity. The angle between the two column vectors in the design has improved from 10.31° to 15.92°. The corresponding $R^2(\mathbf{y}, \widehat{\mathbf{y}}(\lambda)) = 0.3169$ indicates that 97.5% for the squared correlation has been retained. For Step 2, we solve for $\lambda_2 = 0.3898$ to raise $\mathbf{x}_2 \rightarrow \widetilde{\mathbf{x}}_2$. This is the final raised design $\widetilde{\mathbf{X}} = [\widetilde{\mathbf{x}}_1, \widetilde{\mathbf{x}}_2]$. With these values, the squared lengths $\widehat{\boldsymbol{\beta}}'\widehat{\boldsymbol{\beta}} = 4.15$, $\kappa = 27.43$ and $VIF = 7.36$ all showing an improvement in collinearity. The angle between the two column vectors in the design has improved to 21.62°. The corresponding $R^2(\mathbf{y}, \widehat{\mathbf{y}}(\lambda)) = 0.3147$ indicates that 96.9% for the squared correlation has been retained. Row 8 records *MAD* which is 0.009158 by construction.

The values in Column 3 of Table 2 are comparable to the values for the surrogate model from Table 1. However, following Marquardt (1970, p. 610), *VIF* should be less than 10 and thus, for this example, we would favor the raise estimators as the ridge-type method to be used.

Funding
The authors received no direct funding for this research.

Author details
Diarmuid O'Driscoll[1]
E-mail: diarmuid.odriscoll@mic.ul.ie
Donald E. Ramirez[2]
E-mail: der@virginia.edu
[1] Department of Mathematics and Computer Studies, Mary Immaculate College, Limerick, Ireland.
[2] Department of Mathematics, University of Virginia, Charlottesville, VA, USA.

Citation information
Cite this article as: Mitigating collinearity in linear regression models using ridge, surrogate and raised estimators, Diarmuid O'Driscoll & Donald E. Ramirez, *Cogent Mathematics* (2016), 3: 1144697.

References

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

Garcia, C. B., Garcia, J., López Martin, M. M., & Salmeron, R. (2015). Collinearity: Revisiting the variance inflation factor in ridge regression. *Journal of Applied Statistics, 42*, 648–661.

Garcia, C. B., Garcia, J., & 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., & Ramirez, D. (in press). *The successive raising estimator and its relation with the ridge estimator* (under review).

Hadi, A. S. (2011). Ridge and surrogate ridge regressions. In M. Lovric (Ed.), *International encyclopedia of statistical science* (pp. 1232–1234). Berlin: Springer.

Hadi, A. S., & Ling, R. F. (1998). Some cautionary notes on the use of principal component regression. *The American Statistical Association, 52*, 15–19.

Hoerl, A. E. (1964). Ridge analysis. *Chemical Engineering Progress, Symposium Series, 60*, 67–77.

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

Jensen, D. R., & Ramirez, D. E. (2008). Anomalies in the foundations of ridge regression. *International Statistical Review, 76*, 89–105.

Jensen, D. R., & Ramirez, D. E. (2010a). Surrogate models in ill-conditioned systems. *Journal of Statistical Planning and Inference, 140*, 2069–2077.

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

Jensen, D. R., & Ramirez, D. E. (2013). Revision: Variance inflation in regression. *Advances in Decision Sciences,* 1–15. 671204.

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

Marquardt, D. W. (1963). An algorithm for least-squares estimation for nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics, 11*, 431–441.

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. *The American Statistical Association, 29*, 3–20.

McDonald, G. C. (2009). ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics, 1*, 93–100.

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

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

Piegorsch, W., & Casella, G. (1989). The early use of matrix diagonal increments in statistical problems. *SIAM Review, 31*, 428–434.

Riley, J. (1955). Solving Systems of linear equations with a positive definite, symmetric but possibly ill-conditioned matrix. *Math Tables Mathematical Tables and Other Aids to Computation, 9*, 96–101.

Sardy, S. (2008). On the practice of rescaling covariates. *International Statistical Review, 76*, 285–297.

Woods, L. C. S., Holl, K. L., Oreper, D., Xie, Y., Tsaih, S.-W., & Valdar, W. (2012). Fine-mapping diabetes-related traits including insulin resistance, in heterogeneous stock rats. *Physiological Genomics, 44*, 1013–1026.

***Cogent Mathematics* (ISSN: 2331-1835) is published by Cogent OA, part of Taylor & Francis Group.**