The Optimal Method for Picking a Good Ridge Parameter

Scott Duong

STA315

University of Toronto Mississauga

April 10, 2023

Table of Contents

Page

The Optimal Method for Picking a Good Ridge Parameter

Scott Duong

April 10, 2023

1 Abstract

Ridge regression is an important method that is used in data analysis when data suffers from multicollinearity. The ridge parameter is the aspect of ridge regression that controls the shrinkage effect on a model. There are multiple methods for obtaining the ridge parameter. Generalized cross-validation is one of the best methods for selecting the ridge parameter. There are numerous advantages of using this technique over methods such as PRESS and Range Risk Estimate. The inefficiencies of each prominent method of obtaining the ridge parameter will be assessed using MSE, and the best procedure will be determined through the lowest inefficiency score in the simulation.

2 Introduction

In machine learning and data science, various techniques and methods are employed for specific purposes. One of the methods that will be discussed in this paper is regression. Regression is a technique for investigating the relationship between a dependent variable and an outcome. In this paper, the focus will be on the technique of ridge regression.

Ridge regression is a popular technique used in machine learning and statistics. It is a linear regression method used when the data is affected by multicollinearity. This is when several predictor variables in a model are correlated. It differs from linear regression by not using ordinary least squares to estimate parameters; instead, it is biased and utilizes a ridge estimator and ridge parameter. There is significance in choosing the appropriate value for the ridge parameter. There are distinct effects on the model when the ridge parameter is large or small. Therefore, an appropriate method must be selected and applied to determine the value of the ridge parameter. The method in the paper that the authors deemed sufficient is the generalized cross-validation method. A method that is described to have various applications in regression and other fields (Golub et al., 1979). This paper will be a review of the proposition posed by Golub et al (1979).

The rest of the paper is organized as follows. In Section 3, we provide a brief overview of ridge regression and the ridge parameter. In Section 4, we introduce various methods to calculate the ridge parameter. In Section 5, we explain the disadvantages of the other methods and explain the authors' preference for the GCV. Section 6 will deal with a simulated data set and the MSE of four methods will be measured, and the inefficiencies will be compared. Finally, Section 7 discusses the results and provides some concluding remarks.

3 Review

This section will contain a review of ridge regression, which includes various concepts, methods and examples to help facilitate a deeper understanding of ridge regression.

Ridge regression is a type of linear regression applied when a researcher wants to estimate the coefficients of a model, but the predictor variables are correlated. It differs from ordinary linear regression as it is biased and possesses a ridge estimator and a penalty term λ. The ridge estimator is given by

$$\hat{\beta}(\lambda) = (X^T X + \lambda I)^{-1} X^T Y \tag{1}$$

The ridge penalty term controls the shrinkage effect in a model; shrinkage is when the coefficient estimates in the model are brought close to 0, which helps prevent overfitting or underfitting the data in a model. It is essential to select an appropriate value for the ridge estimator. If the ridge parameter approaches 0, the ridge estimator will become closer to the results of an ordinary least squares estimator. If the parameter approaches ∞, then the ridge estimator will approach 0. In addition to this, as λ increases, the bias increases, and the variance decreases. Thus, it is important when implementing ridge regression to select an appropriate method for estimating the best ridge parameter. A non-optimal value can shift your results to be undesired.

4 Existing Methods

This section will contain a brief overview of the existing methods used to determine a value for the ridge parameter. This will include an explanation of the theory, methods and formulas.

4.1 Allen's PRESS

Allen's PRESS is one of the methods of deriving the ridge parameter shown in the paper, and it does not require σ to estimate λ. It works by minimizing the equation

$$P(\lambda) = \frac{1}{n}\sum_{k=1}^{n}([X\beta^{(k)}(\lambda)]_k - y_k)^2 \tag{2}$$

where $\beta^{(k)}(\lambda)$ is the ridge estimate of β with the kth data point y_k omitted (Golub et al., 1979). The theory behind how this estimator works is if λ is an acceptable value, then the kth component $[X\beta^{(k)}(\lambda]_k$ of $X\beta^{(k)}(\lambda)$ is a good predictor of the kth y value (Golub et al., 1979). When the Sherman-Morrison-Woodbury formula is applied to Allen's PRESS it takes the form of

$$P(\lambda) = \frac{1}{n}||B(\lambda)(I - A))y||^2, \tag{3}$$

where $B(\lambda)$ is a diagonal matrix with jjth entry $\frac{1}{1-a_{jj}(\lambda)}$, with a_{jj} as the jjth entry of $A(\lambda) = X(X^TX + n\lambda I)^{-1}X^T$ (Householder, 1964).

4.2 Range Risk

The Range Risk estimate is another one of the methods discussed in the paper for the estimation of the ridge parameter. Golub et al. (1979) range risk estimate originates from the equation $T(\lambda)$ which is the mean square error in estimating $X\beta$

$$T(\lambda) = \frac{1}{n}||X\beta - X\hat{\beta}(\lambda)||^2 \tag{4}$$

Upon taking the expected value of the equation $T(\lambda)$, the equation takes the form of

$$ET(\lambda) = \frac{1}{n}||(I - A(\lambda))g||^2 + \frac{\sigma^2}{n}TrA^2(\lambda) \tag{5}$$

where g is equal to $X\beta$ (Golub et al., 1979). The range risk estimator is an unbiased estimator $\hat{T}(\lambda)$ of $ET(\lambda)$ for $n > p$ (Golub et al., 1979).

$$\hat{T}(\lambda) = \frac{1}{n}||(I - A(\lambda))y||^2 - \frac{2\sigma^2}{n}Tr(I - A(\lambda)) + \hat{\sigma}^2, \tag{6}$$

where $\hat{\sigma}^2 = \frac{1}{n-p}||(I - X(X^TX)^{-1}X^T)y||^2$. By minimizing $\hat{T}(\lambda)$ we are able to find a value for λ (Golub et al., 1979). This method differs from the previously discussed method by requiring the use of σ in order to estimate λ (Golub et al., 1979).

4.3 Generalized Cross-Validation

In the paper, the ideal method for determining the ridge parameter was using the generalized cross-validation method. This is obtained by minimizing the equation V(λ)

$$V(\lambda) = \frac{\frac{1}{n}||(I - A(\lambda)y||^2}{[\frac{1}{n}Trace(I - A(\lambda))]^2} \tag{7}$$

where

$$A(\lambda) = X(X^T X + n\lambda I)^{-1} X^T \tag{8}$$

The generalized cross-validation method is a rotation-invariant method of the Allen Press, also known as ordinary cross-validation (Golub et al., 1979). The GCV method for estimating λ is derived from using Allen's PRESS on a transformed model

$$\tilde{y} = WU^T Y = WDV^T \beta + WU^T \epsilon = \tilde{X}\beta + WU^T \epsilon \tag{9}$$

The data matrix $\tilde{y} = (\tilde{y}_1, \cdots, \tilde{y}_n)^T$ and the design matrix $\tilde{X} = WDV^T$ is a circulant matrix. $A(\lambda)$ and $\tilde{A}(\lambda)$ have the same eigenvalues, and $\tilde{A}(\lambda) = \tilde{X}(\tilde{X} * \tilde{X} + nI\lambda)^{-1}\tilde{X}*$ (Golub et al., 1979). Applying all of these to $P(\lambda)$ derives the equation for $V(\lambda)$ (Golub et al., 1979).

$$\begin{aligned} V(\lambda) &= \frac{\frac{1}{n}||(I - \tilde{A}(\lambda)\tilde{y}||^2}{[\frac{1}{n}Trace(I - \tilde{A}(\lambda))]^2} \\ &= \frac{1}{n}\sum_{v=1}^{n}(\frac{n\lambda}{\lambda_{vn} + n\lambda})^2 z_v^2 \Big/ \left[\frac{1}{n}\sum_{v=1}^{p}\frac{n\lambda}{\lambda_{vn} + n\lambda} + n - p\right]^2 \end{aligned} \tag{10}$$

It can also be shown that GCV is a weighted version of Allen's PRESS

$$V(\lambda) = \frac{1}{n}\sum_{k=1}^{n}([X\beta^k(\lambda)]_k - y_k)^2 w_k^{(\lambda)} \tag{11}$$

where $w_k^{(\lambda)} = \frac{1 - a_{kk}(\lambda)}{1 - \frac{1}{n}TrA(\lambda)}$. Upon using the GCV method, λ can be estimated by minimizing the equations (7) or (10) (Golub et al., 1979).

5 The preference of GCV

This section will contain an overview of the benefits of estimation using GCV, and explain the downsides of the previously described methods.

5.1 Downsides of other methods

When using the other methods present in the paper, there are some clear downsides to employing them. In using, Allen's PRESS, there is an extreme case where λ can not be estimated. Golub et al. (1979) propose a case, where the entries of the matrix X are equal to 0, besides the values for x_{ii}, where $i = 1, 2, \cdots, p$. The matrix $A(\lambda)$ becomes diagonal and the equation $P(\lambda)$ takes the form

$$P(\lambda) = \frac{1}{n}\sum_{k=1}^{n} y_k^2 \tag{12}$$

In this form, we see that Allen's PRESS does not possess a unique minimum anymore, and $[X\beta^k(\lambda)]_k$ is unable to be a predictor value for y_k, and this extends to the near diagonal case (Golub et al., 1979). Under these conditions, a researcher would be unable to confidently estimate a "good" value for λ. So in this case, GCV would perform better than Allen's PRESS.

For the range risk estimate, there is a need to estimate σ to use the method. The need to estimate σ means that it can not be used to solve ill-posed linear operator equations numerically (Golub et al., 1979). Regarding these problems, there is typically no way to estimate a value for σ (Golub et al., 1979). So in this case, a method such as range risk would not be applicable.

5.2 Advantages of GCV

Several prominent methods exist for determining a suitable value for the ridge parameter. However, the best approach as proposed in the paper is Generalized Cross-Validation. When using the GCV method, researchers do not have to estimate a value for σ. An advantage to this unique property is that where problems have $n-p$ equalling a small quantity or when the "real" model takes the form of

$$y_i = \sum_{j=1}^{\infty} x_{ij}\beta_j + \epsilon_i, \quad i = 1, 2, \cdots, n \tag{13}$$

it can still be solved by GCV (Golub et al., 1979). The GCV estimator behaves like a risk improvement estimator, however, it does not require an estimate of σ, so it can be used when $n-p$ is small, or even if $p \geq n$ in certain cases (Golub et al., 1979). Furthermore, as mentioned previously when solving ill-imposed linear operator equations there is no way of estimating σ^2 (Golub et al., 1979). GCV does not require this value, so it would be an applicable method to solve these types of equations. The GCV method has further applications as well in the field of statistics. It is used in subset selection, singular value truncation methods for regression, choosing mixtures among methods, and choosing the order of an auto-regressive model to fit a stationary time series (Golub et al., 1979). It has more advanced applications as well in other fields, such as curve smoothing, numerical differentiation, and the optimal smoothing of density and spectral density estimates (Golub et al., 1979).

6 Simulation

In this section, there will be a simulation that will test each method of generating the ridge parameter to see which performs the best. The code will reproduce the data matrix in the paper, however, it will differ in the method used to test each method. The four methods that will be tested are Generalized Cross-Validation (GCV), Allen's PRESS (APRESS), Range-Risk Estimate (RR), and Maximum Likelihood Estimates (MLE). The data matrix was created with the following parameters for the code to be reproducible. The parameters n and p correspond to the dimensions of the data matrix. So the numbers used in the simulation were from a 21 x 10 size matrix. The ratio of the highest value to the lowest is by a factor of 1.54 x 10^5. The value of $||X\beta||^2 = 370.84$, the euclidean norm squared of $X\beta$. The S/N value is the signal-to-noise ratio and can be calculated by $S/N{=}[\frac{1}{n}\frac{||X\beta||^2}{\sigma^2}]^{1/2}$. The value of S/N in the code is 4200. The simulation was performed over 100 Monte Carlo runs, and the table was organized from lowest inefficiency to highest. To gauge how effective each method is at generating a ridge parameter, the mean square error will be measured to assess the inefficiency of each method. This differs from the paper's method which used two different equations to gauge the inefficiency of each method. Golub et al. (1979) used two equations denoted I_D and I_R given by:

$$I_D = \frac{||\beta - \hat{\beta}_{\hat{\lambda}}||^2}{(min_\lambda||\beta - \hat{beta}_{\hat{\lambda}}||^2)} \tag{14}$$

$$I_R = \frac{T(\hat{\lambda})}{min_\lambda T(\hat{\lambda})} \tag{15}$$

Assessing the inefficiencies of each method based on their MSE score, we arrive at the following conclusion:

Method	Average Inefficiency
GCV	0.08585
MLE	0.08585
APRESS	0.08591
RR	0.08611

As seen in the table, GCV performed the best out of the other methods as it had the lowest value for observed inefficiencies. In this setting, GCV is the best out of the 4 methods presented. Golub et al. (1979) under their settings and method had differing results with the range risk method performing the best and GCV as a close second.

7 Conclusion

As discussed previously in the paper, when applying ridge regression it is essential to select an appropriate method for determining the value of the ridge parameter. The ridge parameter regulates the shrinkage feature in ridge regression, which is the part that "shrinks" the regression coefficients towards 0. If the value is high, then the bias increases and the variance decreases. Three main methods were examined in the paper; Allen's PRESS, range risk, and the GCV method. Allen's PRESS is one of the few methods that did not require an estimate for σ. It's a popular method used to estimate the value of λ, however, it fails in a specific case. In the edge case, where the matrix $A(\lambda)$ becomes diagonal, then λ can not be reliably estimated using this method. Range risk is a method that requires the value of σ to estimate the value of the ridge parameter. Including σ helps improve the estimation of the value of λ. However, this estimation fails when dealing with ill-posed linear operator equations. As σ can not be estimated when dealing with these equations. The method of GCV does not require the value of σ and does not fail where Allen's PRESS does. The GCV is a rotation-invariant form of Allen's PRESS and this allows it to estimate λ under the case where $A(\lambda)$ is diagonal. GCV is a versatile method with other applications in regression and uses in other fields as well. In the simulation presented in the paper, it was shown that GCV performed the best out of the other existing methods. The GCV had the lowest average inefficiency score over 100 Monte Carlo runs. Upon consideration of this and the other advantages of utilizing GCV to estimate the ridge parameter. We can conclude that Generalized Cross-Validation is a suitable method for estimating the ridge parameter.

References

[1] Gene H. Golub, Michael Heath & Grace Wahba (1979) Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter, Technometrics, 21:2, 215-223, DOI: 10.1080/00401706.1979.10489751

[2] Householder, A (1964) *The Theory of Matrices in Numerical Analysis.* Blaisdell.