International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2018): 7.426

Ridge and Lasso Regression for Undergraduate Research with Simulation in R

Di Gao[1], Xiyuan Liu[2], Stephen Scariano[3]

[1, 3] Department of Mathematics and Statistics, Sam Houston State University,1905 University Ave, Huntsville, TX 77340, United States

[2]Department of Statistics and Data Science, University of Central Florida, 4000 Central Florida Blvd, Orlando, FL 32826, United States

Abstract: ***A key part of the undergraduate statistics curriculum is modeling and prediction. Indeed, regression analysis is one of the most frequently used statistical methodsin the sciences. Most introductory statistics textbookscontain one or more chapters discussing regression methodology. However, with the advent of "big data" problems, the classical discussion of introductory regression analysis may not fully capture the current state of the art. This article presents an undergraduate perspective for regression analysis with regulation. Our discussion concentrates on the methodologywhich is then followed by an informative simulation study.***

1. Introduction

Regression analysis is a statistical methodology for estimating the relationships between a dependent variable and one or more independent variables. In simple linear regression, the goal to use one straight line to describe a potential linear relationship between a dependent variable and a single independent predictor; see Figure 1.

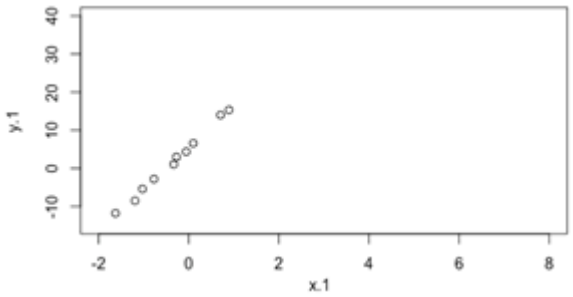

Figure 1

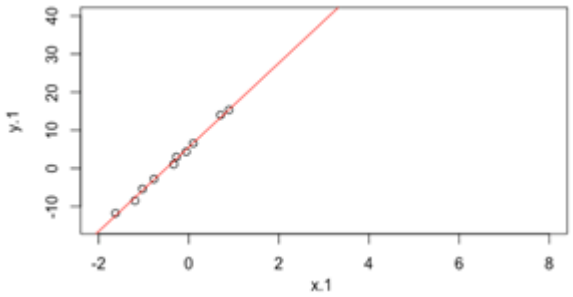

Figure 2

Table 1: Training data

X	-0.27	-1.63	-1.03	-0.33	-1.19	-0.77	0.10	0.89	0.70	-0.06
y	3.00	-11.78	-5.40	1.04	-8.50	-2.80	6.58	15.32	14.01	4.36

Here, the simple statistical model is $y = \beta_0 + \beta_1 x + \varepsilon$, where x represents an independent (explanatory) variable, y denotes a response at x, β_0 and β_1 represents unknown coefficients to be estimated on the basis of data, and ε denotes potential random error. The classical method for estimating the coefficients is due to Laplace who developed the Least SquaresEstimation. It proposes finding estimates $\widehat{\beta_0}$ and $\widehat{\beta_1}$ of the coefficients to minimize $SSE = \sum(y - \widehat{\beta_0} - \widehat{\beta_1}x)^2$, which is called the sum of squarederror (*SSE*). Theestimated least squares lineis then $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}x$,where $\hat{y}$ is the estimated mean response when the explanatory variable is set at x; see the red line in Figure 2.The least square estimators $\widehat{\beta_1} = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}$, $\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$ are statistically unbiased, but sometimes with high prediction variance. This problem of reduced estimation power in prediction in a regression setting is often referred toas *overfitting*in the literature. This situation can directly beseen from thecomplete dataset in Figure 3, where the solid black dots are influential data values.

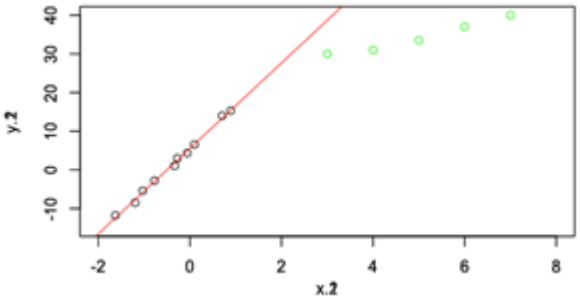

Figure 3

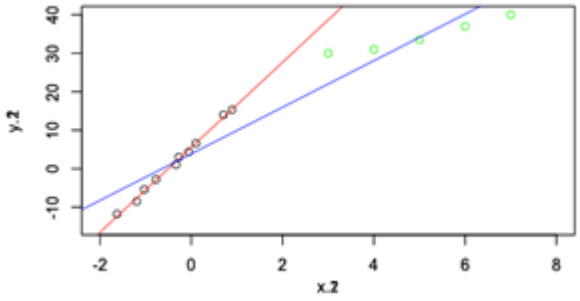

Figure 4

Table 2: Test data

X	3.00	4.00	5.00	6.00	7.00
y	30.00	31.00	33.50	37.00	40.00

The sum of squared prediction error $SSE = \sum(y - \widehat{\beta_0} - \widehat{\beta_1}x)^2$ using the red line in Figure 4 will be very large for the x values corresponding to the solid-colored points there. To solve this issue, the objective the shifts to finding a line with some *bias* in estimating the coefficients that will, nonetheless, produce smaller *SSE*. The blue line in Figure 4

yields lower prediction variance than the red line in Figure 3, and how to achieve this result is discussed next.

2. RIDGE and LASSO Methodologies

With only the limited data information given in Figure 1, we need the ability, specified by a rule, to control the bias and variance trade-off, which is the essence of the concept of *regulation*. In the least squares estimation step, instead of minimizing $\sum(y-\widehat{\beta_0}-\widehat{\beta_1}x)^2$, we instead are interested in minimizing $\sum\,[\,(y-\widehat{\beta_0}-\widehat{\beta_1}x)^2+\lambda\times\widehat{\beta_1}^2]$. The$\widehat{\beta_1}^2$ termis referred to as the L_2 norm, and the corresponding minimization process is called *ridge regression*. The *penalty term*, $\lambda\times\widehat{\beta_1}^2$, forcessome bias into the model. Clearly, the *tuning parameter* λ controls the magnitude of the penalty term, and it is easily seen that as$\lambda\to 0$, then the penalty term also approaches zero. In this event,ridge regression reduces to ordinary least squares regression, so that$\hat{\beta}_{1,Ridge}=\hat{\beta}_1$. On the other hand, as$\lambda\to\infty$, then $\hat{\beta}_{1,Ridge}\to 0$sincethe penalty term dominates in the equation, $\sum\,[\,(y-\widehat{\beta_0}-\widehat{\beta_1}x)^2+\lambda\times\widehat{\beta_1}^2]$. To choose the proper tuning parameter λ, cross validation (i.e., different choices for λ used for comparison) will be undertaken, and the whole process will be demonstrated in asimulation study to follow.
In a real data situation, the information we know is referred to as *training* data, shown as the open-circle dots in Figure 3. The solid black dots in Figure 3 are referred as *test*data. We want to initially fit a model using only the training data, and then consider prediction using the test data as well. For ridge regression, this amounts to minimizing $\sum\,[\,(y-\widehat{\beta_0}-\beta 1x2+\lambda\times\beta 12$in order toshift the slope downward, accounting for the test data geometric configuration.

Let us compare the red and blue fitted linesshown inFigure 4. A modelbased only on the training data set, the open-circle dots, is given by the red line, $\hat{y}=5.54+11.05x$. The ten training data values produce $SSE=3.06$. The fitted equation for blue line is $\hat{y}=3.85+6.05x$ and the sum of squared residuals is $SSE=150.94$. Based on the least squarescriterion, we should select the red line for prediction since it has a smaller sum of squared residuals. However, if we consider ridge regression,the blue line may outperform the red line in the sense of lowing *SSE*. A tuning parameter choice of $\lambda=5$ (a moderate penalty level selection) produces $\sum\,[\,(y-\widehat{\beta_0}-\widehat{\beta_1}x)^2+\lambda\times\widehat{\beta_1}^2=3.06+5\times 11.05^2=613.57$ for the red line and$\sum\,[\,(y-\widehat{\beta_0}-\widehat{\beta_1}x)^2+\lambda\times\widehat{\beta_1}^2=150.94+5\times 6.05^2=333.95$ for the blue line. In this event, the blue line should be our choice produceslower SSE.

A major goal of ridge regression is to avoid overfitting. It is easy tosee that the blue fitted line has performed better in terms of the testing data. For red line, $SSE=\sum((y.test)-\beta 0-\beta 1(x.test)2=4221.79.$ For blue line, $SSE=\sum((y.test)-\widehat{\beta_0}-\widehat{\beta_1}(x.test))^2=121.66$. Clearly, the blue line has significantly lower prediction error. Again, cross-validation should iteratively be used to select the optimal tuning parameter λ.

Lasso regression is similar to ridge regression, except that lasso uses the L_1 norm, $|\widehat{\beta_1}|$, as the penalty term. Everything else remains the same in the minimization process, so the objective is to minimize $\sum\,[\,(y-\widehat{\beta_0}-\widehat{\beta_1}x)^2+\lambda\times|\widehat{\beta_1}|\,]$.Due to Lasso's geometric properties, it can shrink some of the regression coefficients to zero and hence achieve apossible reduction in dimensionality. See Figure 5.

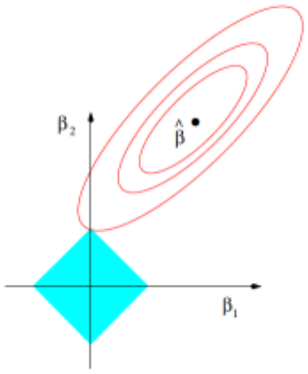

Figure 5

3. Simulationusing Ridge Regression

For ridge regression simulation, the combined training and test datasets given in Tables 1 and 2 are used. Cross-validation is used in order to select the best tuning parameter λ. Usually, a 10-fold cross-validation is applied, but since our data set is very small ($n=15$),we use a 5-fold cross-validation instead. This means that we randomly divide the data into fiveequal subgroups,simply named as groups 1 to 5, each group havingthree observations.

We start with the first group being the test data, and the remaining data are the training data. Since $\lambda\in(0,+\infty)$, we canchoose some reasonable discrete λ ,for example, choosing $\lambda\in(1{,}1000)$ by increments of 1 will usually suffice. When $\lambda=1$, computationally identify the estimated coefficients by minimizing $\sum\,[\,(y-\widehat{\beta_0}-\widehat{\beta_1}x)^2+\lambda\times\widehat{\beta_1}^2$, and record its value as$SSE_1(\lambda=1)$.Following that step, treat the second group as the test data set, and the remaining observations as the training data set. Again, set $\lambda=1$ and find the sum of squared residuals from the testdata set, recorded as$SSE_2(\lambda=1)$.Continue this process until every group is used as a test data set with the remaining data values as the training data set. Finally, stop this process after computing,$SSE_5(\lambda=1)$.Therefore, when$\lambda=1$, the average sum of squared error is $CV(\lambda=1)=\frac{1}{k}\sum_{i=1}^{k}SSE_k(\lambda=1)$, where $k=5$.

The precedingalgorithmis repeated for $\lambda=2,3,\cdots,1000,$ andw e identify $\lambda=\lambda^*$ which minimizes$CV(\lambda)$, say$CV(\lambda^*)$, along with its corresponding estimated coefficients$\widehat{\beta_{0,\lambda^*}}$ and $\widehat{\beta_{1,\lambda^*}}$.

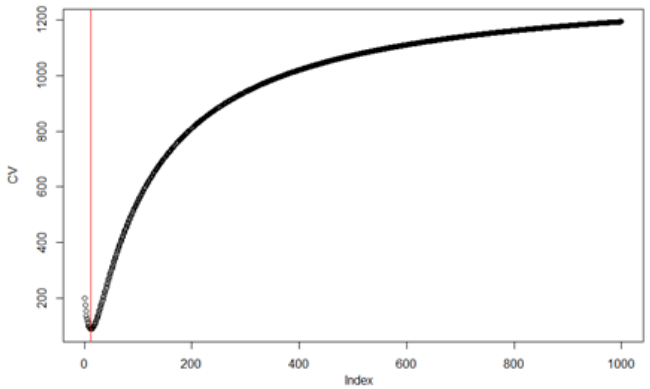

Figure 6: CV values for $\lambda = 1,2,3,\cdots,1000$

Here, as seen in Figure 6,the minimization occurs atthe tuning parameter$\lambda^* = 12$with$CV(\lambda^* = 12) = 89.09$, and the estimated regression coefficients turn out to be $\widehat{\beta_{0,\lambda^*}} = 2.35$ and $\widehat{\beta_{1,\lambda^*}} = 5.79$.

4. Simulationusing Lasso Regression

In the case of six explanatory variables, the model is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + +\varepsilon$ with $SSE = \sum(y - \widehat{\beta_0} - \widehat{\beta_1} x_1 - \widehat{\beta_2} x_2 - \widehat{\beta_3} x_3 - \widehat{\beta_4} x_4 - \beta 5x5 - \beta 6x62$. The R package,**glmnet**,can be used to perform the simulation. To demonstrate dimension reduction, wearbitrarily set theregression coefficients in this model to be those given in Table 3.

Table 3: Possible regression coefficients for a model with six variables

β_0	β_1	β_2	β_3	β_4	β_5	β_6
5	10	15	20	0	0	0

Due to lasso's minimization criterion, the last three coefficients should be reduced. Again, using cross-validation, the best tuning parameter is $\widehat{\lambda_L} = 10.24$, and estimated coefficients for Lasso regression are

Table 4: Estimated regression coefficients for the model

$\widehat{\beta_{L,0}}$	$\widehat{\beta_{L,1}}$	$\widehat{\beta_{L,2}}$	$\widehat{\beta_{L,3}}$	$\widehat{\beta_{L,4}}$	$\widehat{\beta_{L,5}}$	$\widehat{\beta_{L,6}}$
165.46	8.8737	14.6183	19.6056	.	.	.

Lasso regression successfully reduced the dimension while simultaneously retaining the variance reduction properties of ridge regression.

5. Conclusions

The recent advent of powerful computing capabilities now permits both teaching and researchto extend far beyond the frontiers of traditional statistical science. This capability must be valued and pursued not only for undergraduate research opportunities, but also fully integrated into the modern statistics classroom.

References

[1] Anscombe, F.J. (1973). Graphs in Statistical Analysis. The American Statistician, 27(1), 17-21.
[2] Chance, B., & Rossman, A. (2006). Using Simulation to Teach and Learn Statistics. Proceedings of 7th International Conference on Teaching Statistics, Auckland, New Zealand: International Association for Statistical Education.
[3] Hair, J.F. Jr. (2006). Successful Strategies for Teaching Multivariate Statistics. Proceedings of 7th International Conference on Teaching Statistics, Auckland, New Zealand: International Association for Statistical Education.
[4] Tibshirani, R. (1996).Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 58, No. 1, pp. 267-288.
[5] Park, T., Casella, G. (2008). The Bayesian Lasso. Journal of the American Statistical Association. Vol. 103, No. 482, Theory and Methods.
[6] Shen, G., Gao, D., Wen, Q., Magel, R. (2016). Predicting Results of March Madness Using Three Different Methods. Journal of Sports Research. Vol 3, No.1, pp.10-17.
[7] Goldman R. N., & McKenzie, J. D. Jr. (2009). Creating Realistic Data Sets with Specified Properties Via Simulation. Teaching Statistics, 31(1), 7-11.
[8] Friedman, J. (2019). R Package 'glmnet'.