Methods to prevent overfitting and solve ill-posed problems in statistics: Ridge Regression and LASSO

Chris Van Dusen

Advisor: Fred Tinsley

Abstract

Linear regression is one of the most widely used statistical methods available today. It is used by data analysts and students in almost every discipline. However, for the standard ordinary least squares method, there are several strong assumptions made about data that is often not true in real world data sets. This can cause numerous problems in the least squares model. One of the most common issues is a model overfitting the data. Ridge Regression and LASSO are two methods used to create a better and more accurate model. I will discuss how overfitting arises in least squares models and the reasoning for using Ridge Regression and LASSO include analysis of real world example data and compare these methods with OLS and each other to further infer the benefits and drawbacks of each method.

1. Introduction

Consider the standard model of ordinary least squares (OLS) for multiple linear regression

$$Y = X\beta + \epsilon \tag{1}$$

where $y \in \mathbb{R}^n$, $\beta \in \mathbb{R}^p$, and $X \in \mathbb{R}^{nxp}$. We can expand this to $y_i = \sum_{j=1}^{p} \beta_i X_{ij} + \epsilon_i$, $\forall i = 0, 1, ..., n$. Here β_j are non-random unknown parameters, X_{ij} are non-random and observable, and ϵ_i are random so y_i are random. This standard model is used widely across disciplines and is a very powerful tool for any statistician.

1.1. Gauss-Markov Theorem

With the standard model given above, several assumptions are made about the data and model that are not necessarily true for real-world data.

Preprint submitted to Colorado College Department of Mathematics September 16, 2016

The most common assumptions are:

- $E[\epsilon] = 0$
- $V[\epsilon] = \sigma^2$
- $Cov[\epsilon_i, \epsilon_j] = 0 \ \forall \ i \neq j$

If these assumptions are found to be true, the Gauss-Markov theorem states that OLS is the best linear unbiased estimator for the dataset. These assumptions are oftentimes mostly true in smaller datasets, which makes OLS a very powerful tool for statisticians and scientists everywhere. Unfortunately, these assumptions tend to be false with sufficiently large datasets and therefore the OLS method can cause some issues with the resulting model.

1.2. Multicollinearity and Overfitting

One of the most common issues with the OLS method is the tendency for the model to overfit the data when there is too much noise caused by correlated variables. This can happen in many different situations. The most extreme case occurs when $p > n$. From multiple linear regression we have the coefficient estimate

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{2}$$

which we can rewrite as $[(X^T X)^{-1} X^T]^{-1} \hat{\beta} = Y$. We can clearly see that if $p > n$ there exists no unique solution to the system and linear regression fails to produce accurate coefficient values. With less extreme situations multicollinearity can cause the model to be overly sensitive to small changes in parameter values and coefficients can have the "wrong" sign or an incorrect order of magnitude. When a linear model does begin to overfit the data, the coefficients can hive high standard errors and low levels of significance despite a high R^2 value.

1.3. Geometry of Least Squares

Least Squares is most effective when dealing with orthogonal design matrices. When the matrix is ill-conditioned, least squares will still attempt to find a solution such that $d(X\beta, Y)$ is minimized. Presented here is the definition of the least squares solution:

Definition 1. *$\hat{\beta}$ is a least squares solution of the equation system $X\beta = Y$ iff*

$$\forall \beta \in \mathbb{R}^n \; \|Y - X\hat{\beta}\| \leq \|Y - X\beta\|$$

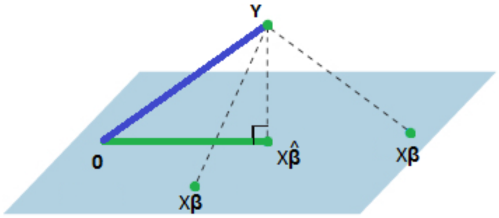


Figure 1: The least squares solution $X\hat{\beta}$ is closer to Y than any other $X\beta$

This figure highlights the importance of orthogonality in the design matrix. With perfectly orthogonal matrix, the least squares solutions will give perfect solutions. As mentioned in section 1.2, multicollinearity can also cause instability in coefficient estimation, in which small changes in parameter values. The figure below shows the problem, and gives some insight on why the method we are studying is called ridge regression:

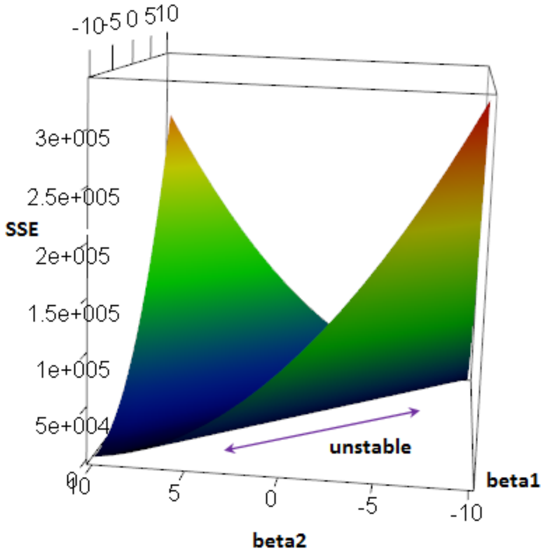


Figure 2: Large changes in the parameter values along the "ridge" of the estimate cause small changes in the prediction error, so estimates are unstable.

2. Ridge Regression and the LASSO

2.1. Introduction to Regularization

Regularization is a method for solving ill-posed problems or problems of models overfitting data. The method involves introducing additional information to a model in the form of a penalty. In terms of Ridge Regression and LASSO, the penalty imposes a shrinkage on the coefficient estimates of ordinary least squares. This penalty controls the instability found in the least squares model with nonorthogonal matrices. Generally, for the L_p regularization term we have $L_p = (\sum_i ||\beta_i||^p)^{\frac{1}{p}}$. Ridge and LASSO deal with the L_2 and L_1 penalties respectively. Regularization is used in preference over other common methods of determining the best linear model, such as best subset selection and stepwise subset selection.

2.2. Ridge Regression

Given the sum of square error estimate for least squares we have $(Y - X\beta)^T(Y - X\beta)$. Ridge regression adds the L_2 penalty such that we have

$$(Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta \tag{3}$$

From equation (3) we can derive the ridge coefficient estimate:

$$
\begin{aligned}
&(Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta \\
&= (Y^T - \beta^T X^T)(Y - X\beta) + \lambda\beta^T\beta \\
&= Y^T Y - Y^T X\beta - \beta^T X^T Y + \beta^T X^T X\beta + \lambda\beta^T\beta \\
&= Y^T Y - 2Y^T X\beta + \beta^T X^T X\beta + \lambda\beta^T\beta \\
&\rightarrow \frac{d}{d\beta} = 0 - 2Y^T X + 2X^T X\beta + 2\lambda\beta = 0 \\
&= \hat{\beta}_{ridge} = (X^T X + \lambda I)^{-1} X^T Y \qquad (4)
\end{aligned}
$$

In equations (3) and (4), $\lambda \geq 0$ is a tuning parameter for the penalty, which is determined separately. When $\lambda = 0$ we have the ordinary least squares estimate, and when $\lambda \rightarrow \infty$ all the coefficients approach zero. The selection of lambda, and thus the optimal model, will be discussed later in the paper. As stated in section 2.1, this penalty shrinks the coefficients. Unlike least squares, which produces only one set of estimates for a model, Ridge Regression produces many sets, depending on what value was assigned to λ. Ridge regression's advantage over ordinary least squares lies in it's bias-variance trade-off. As λ increases, the flexibility of the model fit decreases. This leads to increased bias but decreased variance. There is an existence theorem for ridge regression that states there always exists a $\lambda > 0$ such that the MSE is less than that of the least squares estimate $\lambda = 0$. A proof of the theorem can be found in Hoerl (1970) [1]. However, ridge regression has a major disadvantage to other methods dealing with ill-posed problems and overfitting: it does not perform feature selection. While ridge shrinks coefficients towards zero, the final model chosen will always include all of the predictors (unless $\lambda = \infty$ in which all predictors will be zero). The LASSO is a method that does perform feature selection.

2.3. The LASSO

The LASSO model can be shown in the same form as equation (3) above:

$$(Y - X\beta)^T(Y - X\beta) + \lambda|\beta|_1 \qquad (5)$$

Where $|\beta|_1 = \sum_{j=1}^{p} |\beta|_j$. Comparing equations (3) and (5), we can see that the equations are similar, the only difference is that LASSO uses the L_1 penalty instead of the L_2 penalty. The largest benefit of LASSO is the model's ability to create sparse matrices. The disadvantage that LASSO has from ridge is that because the L_1 penalty contains absolute values, it is much more difficult to solve analytically. Like Ridge, the correct choice of λ and thus the optimal model is very important and will be discussed later in the paper.

2.4. Understanding the Behavior of Ridge Regression and LASSO

In this section we will look at two different formulations of ridge and LASSO to gain better intuition about the methods. First, we will consider the special case such that $n = p$ and X is the identity matrix (1s on the diagonal and 0s elsewhere). We will also assume the intercept of the model is zero. Thus we can write ordinary least squares as $\sum_{j=1}^{p}(y_j - \beta_j)^2$. Here we can easily see the least squares solution is $\beta_j = y_j$. We can find the coefficient estimates of ridge:

$$\begin{gathered}(y_j - \beta_j)^2 + \beta_j^2 \\ = y_j^2 - 2y_j\beta_j + \beta_j^2 + \lambda\beta_j^2 \\ \rightarrow \frac{d}{d\beta} = 0 = -2y_j + 2\beta_j + 2\lambda\beta_j \\ \beta_j = \frac{y_j}{1+\lambda} \qquad (6)\end{gathered}$$

The LASSO estimate is found similarly and thus we obtain:

$$\beta_j = \begin{cases} y_j - \frac{\lambda}{2} & \text{if } y_j > \frac{\lambda}{2} \\ y_j + \frac{\lambda}{2} & \text{if } y_j < -\frac{\lambda}{2} \\ 0 & \text{if } |y_j| \le \frac{\lambda}{2} \end{cases}$$

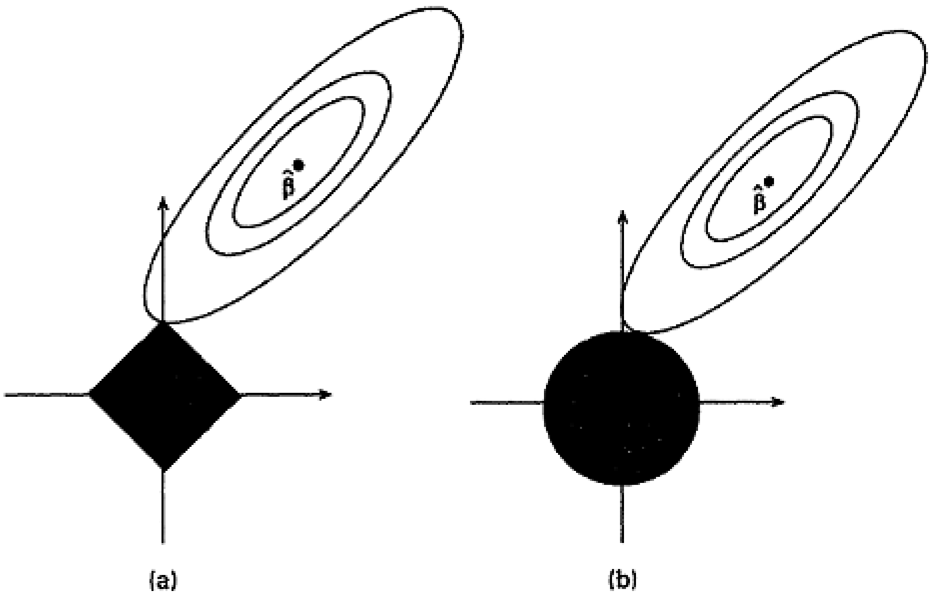
These two estimates show different types of shrinkage. In ridge, the least squares estimate is shrunk at a constant rate for every least squares estimate. LASSO shrinks the least squares estimates at a constant rate unless the least squares estimate absolute value is less than $\frac{\lambda}{2}$; all those coefficients are shrunk to zero

Alternatively, you can show that the Ridge and LASSO models solve these equations respectively:

$$(Y - X\beta)^T(Y - X\beta) \text{ such that } \beta^T\beta \le t \qquad (7)$$

$$(Y - X\beta)^T(Y - X\beta) \text{ such that } |\beta| \le t \qquad (8)$$

This means that for every value of λ in ridge and LASSO, there exists a t such that you will get the same coefficient estimates for (3) and (7), and the same estimates for (5) and (8). When $p = 2$, (7) shows that ridge regression has the smallest RSS out of all the points that lie within the diamond defined by $|\beta_1| + |\beta_2| \leq t$. (8) shows LASSO performs the same with the points that lie within the circle defined by $\beta_1^2 + \beta_2^2 \leq t$. This is illustrated below in figure 3:



Estimation picture for (a) the lasso and (b) ridge regression

Figure 3: Image taken from [5]

In this figure, $\hat{\beta}$ is the least squares solution, and the diamond and circle portray the ridge and LASSO constraints given in (7) and (8). The ellipses around $\hat{\beta}$ are lines of constant RSS. (7) and (8) show that the LASSO and ridge coefficient estimate is where the ellipses and constraint regions meet. Since the constraint region of ridge is a circle, the probability that the intersection will occur on an axis is zero. In contrast, the diamond constraint region has corners at each axis, so the intersection of the ellipses and the constraint region will often occur on the axis.

3. Example data

To study the application of Ridge and LASSO, data was taken from the intergovernmental organization ECMWF's databases available to the public for educational and scientific use. The dataset presented in this paper is fifteen predictors (selected from a preliminary analysis of a set of sixty variables) modeled on a selected response variable. There were 240 observations. Conclusions about the data were not desired, the data was used only to observe how Ridge and LASSO models perform on highly correlated data sets. All computation was done with R software using the package glmnet.

3.1. Standardizing the Predictors

For this dataset, the predictors were standardized. This is because unlike ordinary least squares, which is scale invariant, Ridge and LASSO formulas have the sum of squared coefficients and the sum of absolute value coefficients term respectively. This means that each coefficient estimate is not only dependent on the value of λ but also on the scaling of each predictor. Standardizing the predictors solves any issues with the model that may arise from predictors that have different scales.

3.2. Ridge Trace

When applying Ridge to a set of data, viewing the ridge trace graph can provide valuable insight about the models shrinkage of parameters. the ridge trace plot from our model of the data is given below:

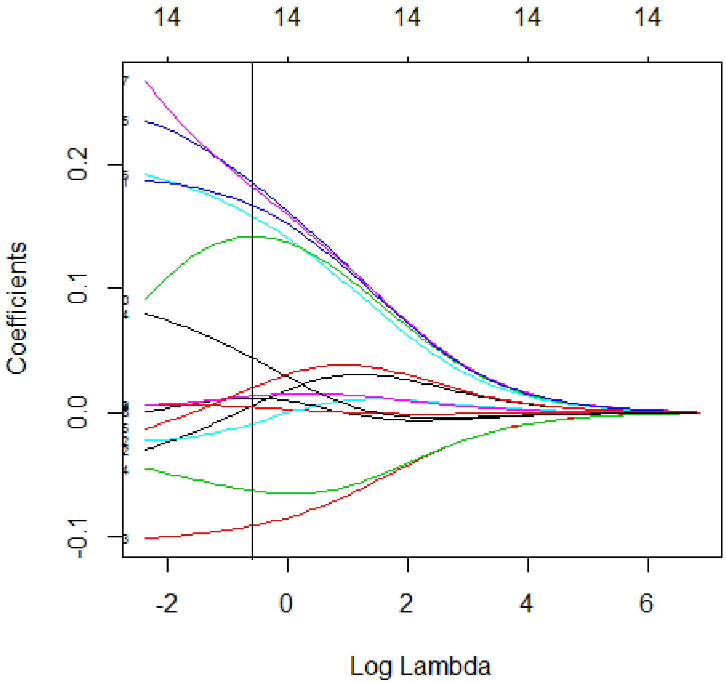


Figure 4: Ridge Trace Plot

In Figure 4, each curve in the plot corresponds to a different ridge coefficient value. These values are plotted against lambda values on the x-axis. The numbers above the graph represent the number of predictors in each lambda model, and is more important when predictor selection occurs in the LASSO model. On the far left of the plot the lambda value is near zero, and thus is representative of the least squares coefficient values. As lambda increases the coefficients shrink towards zero. While in general the coefficients shrink, one can see that some coefficients can infrequently increase as λ increases. The line in the graph represents the model's chosen best value for lambda.

3.3. Lambda selection

The best lambda value of the model is determined from figure 5. In this graph we see the mean-squared error of the model plotted against the λ values in the x-axis. The line shows the λ value that produces the lowest mean square error. The error is determined using cross-validation, which is the most common method for testing models of ridge regression and LASSO. Glmnet performs 10-fold cross validation on the dataset. This means that the data being modelled is split into 10 partitions, and then one partition is chosen as a validation section. The method is then run on the remaining nine partitions and tested against the validation set and a mean square error is determined. This process is repeated so that each partition is used as a validation set and then the mean square errors are averaged and plotted on this graph. This process is used for both Ridge regression and LASSO.

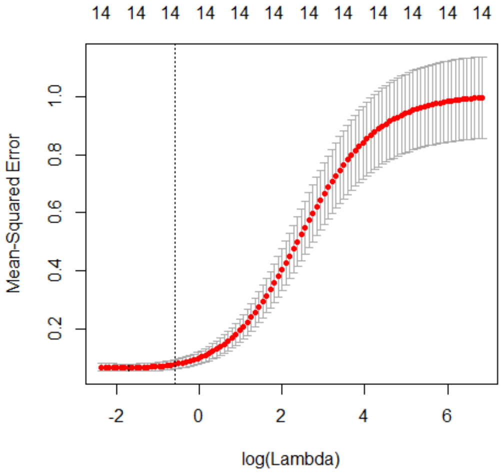


Figure 5: Ridge Cross-Validation Plot

3.4. LASSO Trace

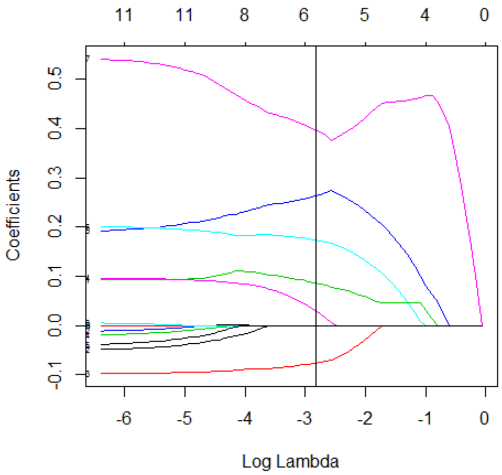


Figure 6: LASSO Trace Plot

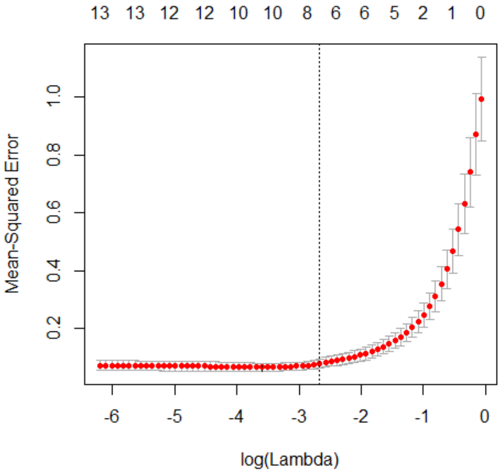


Figure 7: LASSO Cross-Validation Plot

In figures 6 and 7 we see the LASSO equivalent to the ridge trace and cross validation plots shown in sections 3.2 and 3.3. As we can see, the LASSO trace plot exhibits different behavior from the ridge trace, with coefficients disappearing as λ increases. This plot helps visualize the most important predictors in the model very well. As we can see from the selection of the optimal λ value, only 6 predictors are chosen to be used in the final model. The other predictors are sent to zero at a very small λ value and thus have been determined to have no effect on modeling the response variable. The cross-validation plot has a very similar curve for the LASSO as it does for Ridge, which is not surprising.

4. Conclusion

Ridge Regression and LASSO are two methods that improve the overall accuracy of ordinary least squares regression by adding a bias that imposes shrinkage on the model that greatly reduces the variance of coefficient estimates. The methods have been a subject of recent study, and there is still much to learn. This paper reviewed the drawbacks of least squares, discussed how these regularization methods create a better model, and compared the methods. There is still much to learn about the methods, however. Statistical inference on the predictors is still a hot issue that has not been adequately studied, but some good research has been done [6] [7]. A hybrid method of both ridge and LASSO called elastic net has been utilized [3], and while this paper used cross-validation to determine λ the best method for determining an optimal value of lambda is still being researched [4] [8]. Hopefully this paper has encouraged a deeper look into these methods and sparked interest in the use of these methods for big data analysis.

[1] Hoerl A., Kennard R. Ridge Regression: Biased Estimation for Nonorthogonal problems. American Statistics Association 12(1), 55-67 (1970)

[2] Michael Friendly The Generalized Ridge Trace Plot: Visualizing Bias and Precision Journal of Computational and Graphical Statistics (2012)

[3] Art B. Owen A Robust Hybrid of Lasso and Ridge Regression Stanford University, (2006)

[4] B. M. Golam Kibria: Performance of Some New Ridge Regression Estimators Communications in Statistics: Simulations and Computation 32(2), 419-435 (2003)

[5] Robert Tibshirani Regression Shrinkage and Selection via the LASSO Journal of the Royal Statistical Society. Series B (Methodological) 58(1), 267-288 (1996)

[6] Lockhart R., Taylor J., Tibshirani Ry., Tibshirani Ro. A Significance Test for the LASSO Carnegie Mellon University (2015)

[7] Cule E., Vineis P., Iorio M. Significance Testing in Ridge Regression for Genetic Data BMC Bioinformatics 12(372), (2011)

[8] Roa R.B., Fung G, Rosales R. On Dangers of Cross-Validation. An Experimental Evaluation In Proc. SIAM Data Mining (SDM), (2008)

[9] James G., Witten D., Hastie ., Tibshirani R. An Introduction to Statistical Learning Springer Texts in Statistics, New York, (2013).