Linear Model Selection and Regularization

Andrew Andreas

August 15, 2024

Please report any errors to andrew.andreas0@gmail.com

1 Introduction and Motivation

In the previous set of notes we discussed the linear regression model

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p x_p + \epsilon$$

However, we didn't touch on practical issues like how we can improve the *prediction accuracy* of our models and on *model interpretability*.

Prediction Accuracy: In the case where we have abundant data where $n >> p$ then the least squares estimate will have low variance - we can see an example of this where we consider the distribution of the residuals

$$r_i \sim \mathbb{N}\Big(0, \sigma^2\big(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{s_{xx}}\big)\Big)$$

In the case of large n, the variance of the residuals tends towards the true population variance. In the case where n is not much larger than p then there tends to be higher variance, leading to poor generalization capabilities. In the event $n < p$ then we have an undetermined system where least squares does not provide a unique solution

Model Interpretability: When we have a large number of predictors, interpreting the results of a linear regression model can be challenging. In addition, if we are unsure of the relevance of features to a model, it would be useful if we had methods at our disposal which allowed us to automatically perform feature selection so that only relevant features are retained.

In this set of notes I want to outline some of the methods used to deal with these issues.

2 Subset Selection

One of the methods available to address these problems is through subset selection, that is the choice of which regression coefficients we should include in the model

2.1 Best Subset Selection

Best Subset Selection is an exhaustive approach to feature selection where we consider all possible combinations of variables. The general algorithm is given below

Algorithm 1 Best Subset Selection

1. Let $\mathcal{M}_0$ define the null model, the model with no predictors, so that the model just predicts the sample mean for each observation
2. For $l = 1, ..., p$:
   - For all $\binom{p}{l}$ models that contains l predictors, run a k-fold cross validation
   - Pick the model with the best average performance among the k training sets using the RSS or R^2 as the criterion best model among all the models and call this $\mathcal{M}_l$.
3. From the set of models $\mathcal{M}_0, ..., \mathcal{M}_l$ compute the average validation error across the k folds. Use the AIC, BIC or adjusted R^2 as the selection criterion

The issue with this algorithm is that it does not scale well with the number of predictors p with there being 2^p models that need be fit. As such, this method should only be considered for small p.

2.2 Forward Stepwise Selection

An alternative computationally more efficient method is that of Forward Stepwise Selection which begins with the null model and that proceeds to add the predictor which results in the *greatest gain* in performance. The algorithm is detailed below

Algorithm 2 Forward Stepwise Selection

1. Let $\mathcal{M}_0$ define the null model, the model with no predictors
2. For $l = 0, ..., p - 1$:
   - Using k-fold cross validation, consider all $p-l$ models which augments $\mathcal{M}_l$ with a single additional parameter
   - Pick the best model among the $p - l$ possible models based on the either the largest decrease in RSS or greatest increase in R^2 and call this model $\mathcal{M}_{l+1}$
3. From the set of models $\mathcal{M}_0, ..., \mathcal{M}_p$ compute the average validation error across the k folds. Use the AIC, BIC or adjusted R^2 as the selection criterion

This results in $\sum_{i=0}^{p-1}(p - k)$ models which is a substantial decrease from the best subset selection method.

3 Shrinkage Methods

The subset selection methods involve fitting a least squares model for a subset of the predictors. An alternative approach is to fit a model that *constrains* or *regularizes* the coefficient estimates.

3.1 Ridge Regression

In the least squares problem we minimize the RSS to find an optimal solution for the regression coefficients

$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 + \sum_{j=1}^{n} \beta_j x_{ij} \right)^2$$

Ridge regression modifies the RSS by adding a penalization parameter

$$\sum_{i=1}^{n} \left(y_i - \beta_0 + \sum_{j=1}^{n} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$

where $\lambda \geq 0$ is a tuning parameter which controls the level of regularization.

The intuition behind the ridge regression problem is that the RSS term is minimized when our solutions for β_j fit the data well but the penalty term is minimized when the coefficients β_j are small or close to zero and has the effect of *shrinking* the coefficients.

When $\lambda = 0$ we have the least squares solution and as $\lambda \to \infty$ the penalty term dominates and the only way to minimize the expression is when $\beta_j = 0, \forall j = 1, ..., p$. As a result, we generate a unique set of coefficient estimates $\hat{\beta}^R_{\lambda,j}$ for each value of λ which can be done through cross-validation.

3.1.1 Solution

The ridge regression problem does not attempt to shrink the intercept β_0, only the variable coefficients $\beta_1, ..., \beta_p$. If we mean-center the variables

$$\tilde{x}_{ij} = x_{ij} - \bar{x}_j$$

Then we have that $\beta_0 = \bar{y}$. We can estimate the rest of the coefficients using the centred variables. This can be written more concisely using matrix notation

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^T \boldsymbol{\beta}$$

where $\mathbf{X} = \mathbf{X} - \bar{\mathbf{X}} \in \mathbb{R}^{p \times p}$. Like the OLS solution, ridge regression has a closed form solution

$$\hat{\boldsymbol{\beta}} = (\mathbf{X^T X} + \lambda \mathbf{I})^{-1} \mathbf{X^T y}$$

where $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ is a p-dimensional vector that is augmented afterwards with β_0 to yield the $(p+1)$-dimensional solution vector.

3.1.2 Robustness to Collinearity

One of the main problems with simple regression is collinearity, which makes solutions unstable or not uniquely determined.

Collinearity means that one or more of the columns of the matrix $\mathbf{X}$ are correlated. While the OLS solution still exists since $\mathbf{X^T X}$ is non-singular, there is more uncertainty in the solution. When variables are collinear, $\mathbf{X^T X}$ is ill-conditioned, which means that small changes in the input space can cause large changes in the output space.

When the columns of $\mathbf{X}$ are collinear, $(\mathbf{X^T X})^{-1}$ is nearly-singular and the eigenvalues are close to zero. The inverse is given by

$$(\mathbf{X^T X})^{-1} = \frac{1}{\det(\mathbf{X^T X})} \mathrm{adj}(\mathbf{X^T X})$$

Since the determinant of a matrix is the product of its eigenvalues, $\det(\mathbf{X^T X}) = \prod_{i=1}^{p} \lambda_i$, then the matrix $(\mathbf{X^T X})^{-1}$ becomes large. As a result, the variance of the regression coefficients are large in turn, since $\mathbb{V}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X^T X})^{-1}$.

The plausible range of values for β_j is thus much larger and so there is greater uncertainty. A larger variance and standard error results in a smaller t-statistic. This will lead to larger p-values and in turn failing to reject the null hypothesis (e.g., $H_0 : \beta_j = 0$) since $P\big(t(\mathbf{X}) > t(\mathbf{x}) \mid H_0\big) > \alpha$. We can then say that collinearity reduces the power of a test - the probability that we reject the null hypothesis when it is false.

Perfect collinearity means that one or more of the columns of the matrix $\mathbf{X}$ are linearly dependent. This means that $\text{rank}(\mathbf{X}) < p$ and in turn $\text{rank}(\mathbf{X^T X}) = \min\big(\text{rank}(\mathbf{X}), \text{rank}(\mathbf{X}^T)\big) < p$. As a result, $(\mathbf{X^T X})^{-1}$ does not exist.

Ridge regression adds a positive constant to the diagonal of $\mathbf{X^T X}$, making $(\mathbf{X^T X} + \lambda\mathbf{I})$ non-singular, even if $\mathbf{X^T X}$ is singular.

$\mathbf{X^T X}$ is positive semi-definite and we can show that $(\mathbf{X^T X} + \lambda\mathbf{I})$ is positive definite. We can see this using the matrices quadratic form. Consider any vector $\mathbf{a} \neq \mathbf{0}$, then

$$\mathbf{a}^T(\mathbf{X^T X} + \lambda\mathbf{I})\mathbf{a} = \underbrace{\mathbf{a}^T\mathbf{X^T X}\mathbf{a}}_{\geq 0} + \lambda \underbrace{\mathbf{a}^T\mathbf{a}}_{> 0}$$

For any $\lambda > 0$, then $\lambda\mathbf{a^T a} > 0$ and it follows then that $\mathbf{a}^T(\mathbf{X^T X} + \lambda\mathbf{I})\mathbf{a} > 0$ and hence positive definite. If $\lambda = 0$, we have the *OLS* solution which has no unique solution under perfect collinearity.

Since $(\mathbf{X^T X} + \lambda\mathbf{I})$ is positive definite then it has positive eigenvalues, a non-zero determinant and is thus non-singular.

3.1.3 Scaling

Ridge regression does not benefit from the same *scale invariance* that the *OLS* solution does. If we consider data of different scales (e.g. height variable measured in metres and waist size measured in cm), then it is clear that the data are of different scales.

When the data have different initial scales, this is equivalent to scaling the data by some diagonal matrix $\mathbf{D}$ with positive diagonal elements. In the example with height and waist size, $\mathbf{D}$ would take the form

$$\mathbf{D} = \begin{pmatrix} 1 & 0 \\ 0 & 100 \end{pmatrix}$$

since 1m = 100cm.

So when the data have different scales, we essentially have a data matrix $\tilde{\mathbf{X}} = \mathbf{XD}$, where $\mathbf{X}$ would be data of the same scale so that $\mathbf{D} = \mathbf{I}$.

OLS Regression The RSS for linear regression is

$$\begin{aligned} RSS =& (\mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\beta})^T(\mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\beta}) \\ =& (\mathbf{y} - \mathbf{XD}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{XD}\boldsymbol{\beta}) \end{aligned}$$

So that the solution for the vector of regression coefficients is

$$\hat{\boldsymbol{\beta}} = \mathbf{D}^{-1}(\mathbf{X^T X})^{-1}\mathbf{X^T y}$$

where we can see the *OLS* solution is scaled by the inverse of the scaling matrix $\mathbf{D}$. Thus, scaling the data according to some positive matrix $\mathbf{D}$ results in no difference in our estimates for $\mathbf{y}$

$$\begin{aligned} \hat{\mathbf{y}} =& \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}} \\ =& \mathbf{XDD}^{-1}(\mathbf{X^T X})^{-1}\mathbf{X^T y} \\ =& \mathbf{X}(\mathbf{X^T X})^{-1}\mathbf{X^T y} \\ =& \mathbf{X}\hat{\boldsymbol{\beta}} \end{aligned}$$

However, for ridge regression, we see the effect of scaling by considering the solution for $\hat{\boldsymbol{\beta}}$

$$\begin{aligned}\hat{\boldsymbol{\beta}} =&\left(\mathbf{D^T X^T X D} + \lambda \mathbf{I}\right)^{-1} \mathbf{D^T X^T y} \\ =&\Big(\mathbf{D}\big(\mathbf{X}^T\mathbf{X} + \lambda \mathbf{D}^{-2}\big)\mathbf{D}\Big)^{-1} \mathbf{D^T X^T y} \\ =&\mathbf{D}^{-1}\big(\mathbf{X}^T\mathbf{X} + \lambda \mathbf{D}^{-2}\big)^{-1}(\mathbf{D}^{-1}\mathbf{D})\mathbf{X^T y} \\ =&\mathbf{D}^{-1}\big(\mathbf{X}^T\mathbf{X} + \lambda \mathbf{D}^{-2}\big)^{-1}\mathbf{X^T y}\end{aligned}$$

and the prediction vector is given by

$$\begin{aligned}\hat{\mathbf{y}} =&\mathbf{X}(\mathbf{D}\mathbf{D}^{-1})\big(\mathbf{X}^T\mathbf{X} + \lambda \mathbf{D}^{-2}\big)^{-1}\mathbf{X^T y} \\ =&\mathbf{X}\big(\mathbf{X}^T\mathbf{X} + \lambda \mathbf{D}^{-2}\big)^{-1}\mathbf{X^T y}\end{aligned}$$

We can see that λ is proportional to the inverse square of the matrix $\mathbf{D}$. The issue with this formulation is that if $\mathbf{D}$ does not scale each variable by the same factor, it implies we have different values of λ for each variable.

This is problematic for the ridge regression model since if the scale of each β_j is different, they will have non-uniform sensitivity to λ.

This is why **standardization** of the data prior to constructing the solution is necessary. This means that $\mathbf{D} = \mathbf{I}$, so each β_j is scaled by the same λ and thus have the same sensitivity to the regularization parameter.

3.2 Lasso Regression

Another form of regularization replaces the ridge penalty with an L_1 lasso penalty $\sum_{j=1}^{p} |\beta_j|$. The formulation is identical to the ridge case except for the penalty term

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j|$$

Since the absolute value function is not differentiable over its entire domain (specifically at 0), there is not a closed-form solution. However, there are efficient numerical algorithms that can solve the problem.

Despite the lack of closed-form solution, we follow a similar procedure for the data. We mean centre the data so that the intercept $\beta_0 = \bar{y}$.

3.2.1 L_1 Regularization

Like the L_2 regularization in ridge, the L_1 penalty shrinks the values of β_j to minimize the target function. However, the L_1 penalty shrinks some of the variables to zero thereby performing *variable selection*. This is a key difference between the ridge and lasso regression models.

The benefit of lasso is that we get a method which automatically performs subset selection which is computationally cheaper than running subset selection methods. An additional benefit is the increased model interpretability compared to ridge regression. Ridge regression minimizes the values of the coefficients β_j but will not force them to zero leaving us with a model which uses all variables, even if some are very small. Lasso on the other hand does force coefficients to zero and thus results in a *sparse model* - one which only a subset of the variables are used and is thus easier to interpret.

3.2.2 Differences in Regularization

The differences between ridge and lasso regression lie in the regularization used. This can be more easily understood by reformulating both into constrained optimization problems.

Ridge regression can be formulated as

$$\arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 + \sum_{j=1}^{n} \beta_j x_{ij} \right)^2$$
$$\text{subject to} \quad \sum_{j=1}^{p} \beta_J^2 \leq s$$

Lasso regression can be formulated as

$$\arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 + \sum_{j=1}^{n} \beta_j x_{ij} \right)^2$$
$$\text{subject to} \quad \sum_{j=1}^{p} |\beta_J| \leq s$$

If we consider a simple case with two independent variables, then it allows us to visualise the RSS function subjected to the constraints.

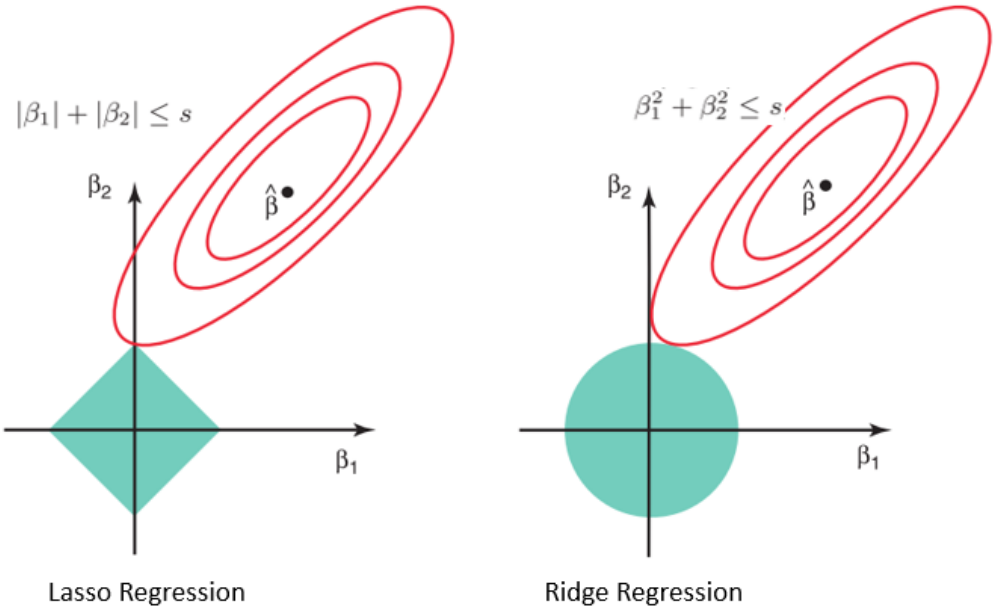

Figure 1: RSS contours plotted with the lasso and ridge constraint regions

Since the lasso constraint has corners at the axes the contour line of the RSS is likely to intersect at an axis. This is what forces some of the coefficients to zero. Since the ridge constraint region is smooth, this is why the coefficients are non-zero - intersection does not occur at the axis.

3.3 Ridge vs Lasso

There are benefits and drawbacks to the use of ridge or lasso. Generally ridge regression may be the preferred solution when

- You wish to retain all the parameters in the model
- There is collinearity amongst the features. As we saw, ridge produces a stable solution even in the presence of perfect collinearity. In addition, since collinear features can contain useful additional information, by retaining all variables, we don't lose potentially informative data

With highly collinear variables, they are nearly perfectly collinear and thus contain very similar information. Suppose we are trying to predict Y and we have standardized predictor variables x and z that are highly collinear. If we take three possible linear combinations of the variables that have roughly equivalent predictive power since $x \approx z$

- $0.2x + 0.8z$,
  - **Ridge penalty**: $\sum_{j=1}^{2} \beta_j^2 = 0.2^2 + 0.8^2 = 0.68$
  - **Lasso penalty**: $\sum_{j=1}^{2} |\beta_j| = |0.2| + |0.8| = 1$
- $0.3x + 0.7z$,
  - **Ridge penalty**: $\sum_{j=1}^{2} \beta_j^2 = 0.3^2 + 0.7^2 = 0.58$
  - **Lasso penalty**: $\sum_{j=1}^{2} |\beta_j| = |0.3| + |0.7| = 1$
- $0.5x + 0.5z$
  - **Ridge penalty**: $\sum_{j=1}^{2} \beta_j^2 = 0.5^2 + 0.5^2 = 0.50$
  - **Lasso penalty**: $\sum_{j=1}^{2} |\beta_j| = |0.5| + |0.5| = 1$

In the case of lasso, the penalty is 1 in all cases and so the coefficient choice does not matter when reducing error. Whereas in ridge, equally weighted coefficients minimizes the penalty term.

The benefits of using lasso are

- When we care a lot about interpretability. The sparse model will be easier to interpret
- If we have a large dataset that we suspect has many redundant features. Many of these will be forced to zero