Contents

[Ok, What is over-fitting? 1](#_Toc522029836)

[How to solve this Over-fitting Problem? 2](#_Toc522029837)

[What’s a cost function? 2](#_Toc522029838)

[What is regularization 3](#_Toc522029839)

[Why penalize the magnitude of coeffcients 4](#_Toc522029840)

[Ridge regression 4](#_Toc522029841)

[Lasso regression 5](#_Toc522029842)

[Elastic net regression 6](#_Toc522029843)

[How to find lambda 7](#_Toc522029844)

[Why feature scaling? 7](#_Toc522029845)

[Explain the difference between L1 and L2 regularization methods. 7](#_Toc522029846)

[When to use regularization methods for regression? 7](#_Toc522029847)

In general, if the relationship between the response (Y) and the predictors (X) is approximately linear, the least squares estimates will have low bias. If the number of observations (N) is much larger than the number of predictors (sometimes denoted as p/P), then the least squares estimates tend to also have low variance, and hence will perform really good on the test set with low error and high R2. On the other hand, if N is not much larger than P, then there can be a lot of variability in the least squares fit, resulting in over-fitting and therefore resulting in poor predictions. And if P > N, then there is no longer a unique least squares coefficient estimate: the variance is infinite so the method cannot be used at all.

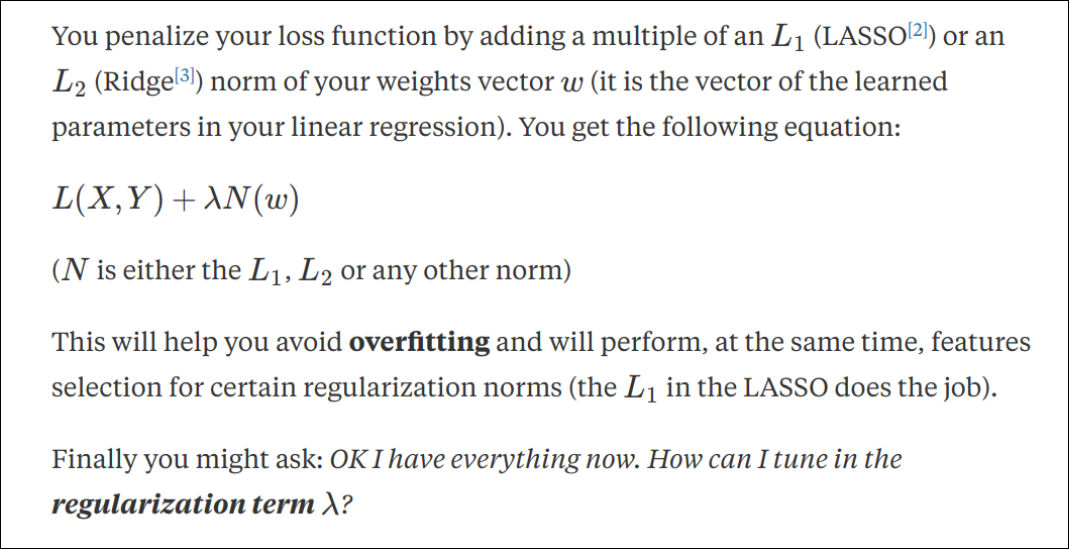
That is an important limitation with OLS. One way to solve this problem is to reduce/remove predictors that are not so important in the model – this is called shrinking. By shrinking the estimated coefficients, we can often substantially reduce the variance at the cost of a negligible (depending on the regularization parameter value) increase in bias. This can lead to substantial improvements in the accuracy with which we can predict the response for observations not used in model training.

# Ok, What is over-fitting?

What is this problem of overfitting and why it arises? Let’s take a sample problem – assume that you want to predict the height of person based on his/her age using a linear regression model. Your response variable (Y) is height and the predictor or independent variable (X) is age. What do you think how the model will perform? Not so good, right? Well, it’s too simple.

Next – you have additional variables that you can add to the model – weight, sex, location. Well, what you did here is add complexity to your data and might have increased the prediction accuracy of the model. Now, you add more variables to your model – height of parents, profession of parents, social background, number of children, weight, number of books, preferred color, best meal, last holidays destination and so on and so forth. That’s one too many variables – and mostly not all of them can explain someone’s height. Your model might do good but it is probably overfitting, i.e. it will probably have poor prediction and generalization power: it sticks too much to the training data and the model has probably learned the background noise while being fit. When tried on an unseen data, this model will perform poorly.

# How to solve this Over-fitting Problem?



It is here where the regularization technique comes in handy. There are two main techniques used with linear regression (L1 or Lasso) and (L2 or Ridge). In general, a regularization term is introduced to a loss/cost function.

# What’s a cost function?

Whenever a model is trained on a training data and is used to predict values on a testing set, there exists a difference between the true and predicted values. The closer the predicted values to their corresponding real values, the better the model. That means, a cost function is used to measure how close the predicted values are to their corresponding real values. The function can be minimized or maximized, given the situation/problem. For example, in case of ordinary least squares (OLS), the cost f unction would be:

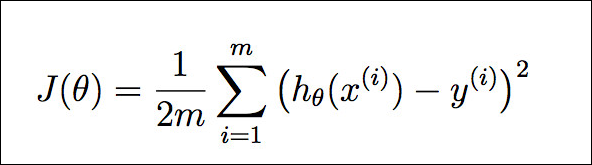
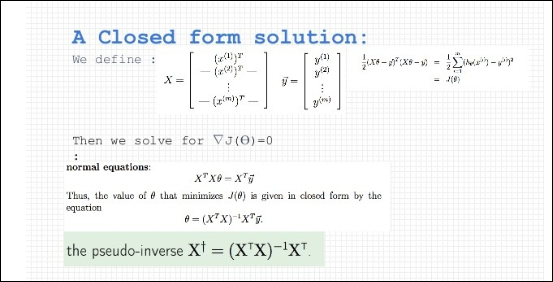


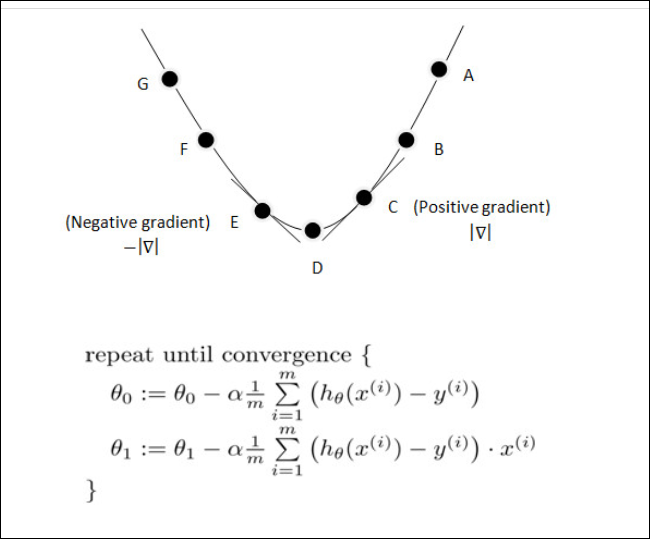
Fig 2. Cost function for OLS

where, J denotes the cost function, m is the number of observations in the dataset, h(x) is the predicted value of the response and y is the true value of the response. In case of OLS, the goal is to minimize this. There are two ways to solve OLS – one using closed form solution, requires doing matrix inverse, multiplication etc. (shown in fig 3) and the other using gradient descent and cost function (fig 4).

**Closed form**



**Gradient Descent form**



# What is regularization

The number of parameters taken in the model for training using observed data is more than the required numbers to represent the problem, which helps to generalize the problem well. But for remedy of over-fitting regularization techniques are added along with the parameters. In the problem that was discussed above, not all variables are required to predict the height of a person. What regularization does is gives more importance to only important parameters and ignore others, thus reducing the complexity of the model.

Lasso and Ridge regressions are closely related to each other and they are called shrinkage methods. We use Lasso and Ridge regression when we have a huge number of variables in the dataset and when the variables are highly correlated.

For both ridge and lasso you have to set a so-called "meta-parameter" that defines how aggressive regularization is performed. Meta-parameters are usually chosen by cross-validation. For Ridge regression the meta-parameter is often called "alpha" or "L2"; it simply defines regularization strength. For LASSO the meta-parameter is often called "lambda", or "L1". In contrast to Ridge, the LASSO regularization will actually set less-important predictors to 0 and help you with choosing the predictors that can be left out of the model. The two methods are combined in "Elastic Net" Regularization. Here, both parameters can be set, with "L2" defining regularization strength and "L1" the desired sparseness of results.

# Why penalize the magnitude of coeffcients

size of coefficients increase exponentially with increase in model complexity

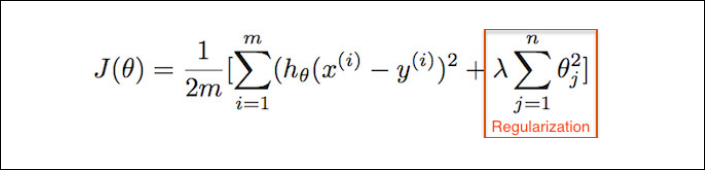
# Ridge regression/L2 regularsisation

Ridge regression is often used when the independent variables are colinear. One issue with colinearity is that the variance of the parameter estimate is huge. Ridge regression reduces this variance at the price of introducing bias to the estimates.

**When to use ridge**

The mean squared error is a measure of the quality of an estimator and is defined as the sum of the variance plus the square of the bias. Often we focus on unbiased estimates but, in some situations, the unbiased estimate can result in very large variances and, as a result, large MSEs.

Ridge regression is a method that seeks to reduce the MSE by adding some bias and, at the same time, reducing the variance.

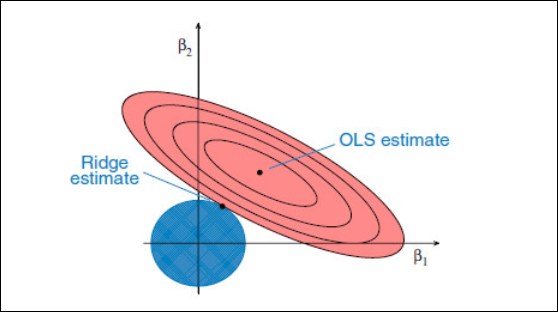


In addition to the cost function we had in case of OLS, there is an additional term added (in red), which is the regularization term. θ is the norm of the coefficients and for ridge regression, the addition is of λ (the regularization parameter) and θ^2 (norm of coefficient squared). The addition of regularization term penalizes big coefficients and tries to minimize them to zero, although not making them exactly to zero. This means that if the θ’s take on large values, the optimization function is penalized. We would prefer to take smaller θ’s, or θ’s that are close to zero to drive the penalty term small. It is also called L2 regularization and it pushes weight with force vectors perpendicular to the surface of a sphere, so they’re likely to be pretty similar, since most of the volume of the sphere lies in areas where weights are similar.

It seeks to reduce the MSE by adding some bias and, at the same time, reducing the variance. Remember high variance correlates to a over-fitting model.

**When to use Ridge Regression?**

When there are many predictors (with some col-linearity among them) in the dataset and not all of them have the same predicting power, L2 regression can be used to estimate the predictor importance and penalize predictors that are not important. One issue with co-linearity is that the variance of the parameter estimate is huge. In cases where the number of features are greater than the number of observations, the matrix used in the OLS may not be invertible but Ridge Regression enables this matrix to be inverted.



Consider a simple example where the number of predictors is 2, the eclipses in the red is the contour plot of the OLS and the circle. The solution for this is the black spot at the center of the red eclipses, which is also the minimum of the function.

Another contour plot in blue is that of the regularized term, denoted by λ(θ1^2+θ2^2). In case of ridge regression, the objective is to reduce the sum of these two values which will come when these two contours meet.

The larger the penalty (λ), the “more narrow” blue contours we get, and then the plots meet each other in a point closer to zero.

The smaller the penalty (λ), the contours expand, and the intersection of blue and red plots comes closer to the center of the red circle (non-penalized solution).

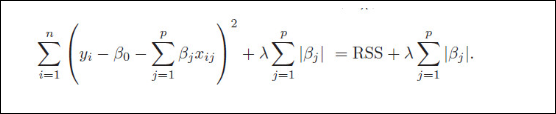
# Lasso regression

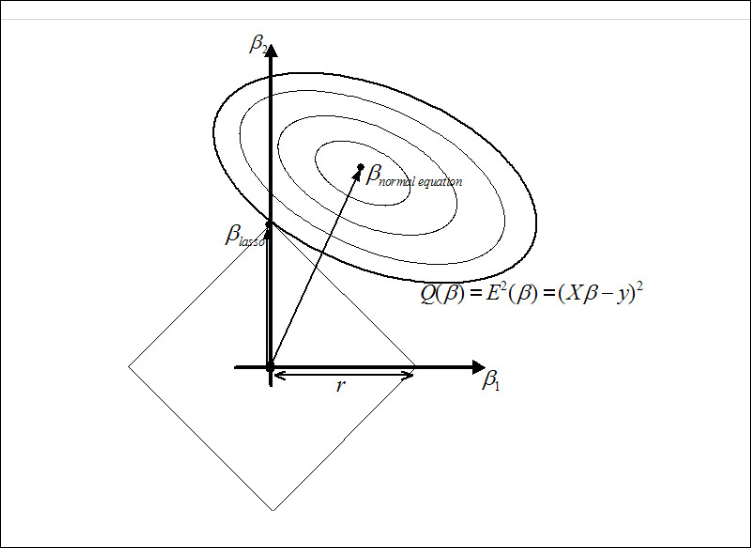
* This is used in logistic regression for feature selection
* Performs L1 regularization, i.e. adds penalty equivalent to **absolute** **value** of the magnitude of coefficients

(beta 1 + beta 2)

One of the things that Ridge can’t be used is variable selection since it retains all the predictors. Lasso on the other hand overcomes this problem by forcing some of the predictors to zero.

Lasso has one small change in the cost equation, instead of the squared of the norm, it takes the absolute value.





The lasso performs shrinkage, so that there are “corners” in the constraint, which in two dimensions corresponds to a rhombus. If the sum of squares “hits” one of these corners, then the coefficient corresponding to the axis is shrunk to zero. As the number of predictors increases, the multidimensional rhombus has an increasing number of corners, and so it is highly likely that some coefficients will be set equal to zero. Hence, the lasso performs shrinkage and (effectively) subset selection. In contrast with subset selection, Lasso performs a soft thresholding: as the smoothing parameter is varied, the sample path of the estimates moves continuously to zero.

# Elastic net regression

Elastic Net produces a regression model that is penalized with both the L1-norm and L2-norm. The consequence of this is to effectively shrink coefficients (like in ridge regression) and to set some coefficients to zero (as in LASSO)

# How to find lambda

Small validation sets are made from training data set and average the performance from the same validation data.

# Why feature scaling?

"The main advantage of scaling is to avoid attributes in greater numeric ranges dominating those in smaller numeric ranges."

# Explain the difference between L1 and L2 regularization methods.

# [When to use regularization methods for regression?](https://stats.stackexchange.com/questions/4272/when-to-use-regularization-methods-for-regression)

Whenever you are facing one of these situations:

large number of variables or low ratio of no. observations to no. variables (including the n≪p case),

high collinearity,

seeking for a sparse solution (i.e., embed feature selection when estimating model parameters), or

accounting for variables grouping in high-dimensional data set.

**What is the difference between lasso and ridge**

Ridge and Lasso regression uses two different penalty functions. Ridge uses l2 where as lasso go with l1. In ridge regression, the penalty is the sum of the squares of the coefficients and for the Lasso, it’s the sum of the absolute values of the coefficients. It’s a shrinkage towards zero using an absolute value (l1 penalty) rather than a sum of squares(l2 penalty).

As we know that ridge regression can’t zero coefficients. Here, you either select all the coefficients or none of them whereas LASSO does both parameter shrinkage and variable selection automatically because it zero out the co-efficients of collinear variables. Here it helps to select the variable(s) out of given n variables while performing lasso regression.

Another type of regularization method is ElasticNet, it is hybrid of lasso and ridge regression both. It is trained with L1 and L2 prior as regularizer. A practical advantage of trading-off between Lasso and Ridge is that, it allows Elastic-Net to inherit some of Ridge’s stability under rotation.

## What is effect of collinearity on Lasso regularization.

LASSO regression can be visualized similarly to ridge regression, but since c is defined by the sum of absolute values of beta, rather than sum of squares, the area it constrains is diamond shaped rather than circular.  Figure 2 shows the selection of the beta estimator from LASSO regression. As you can see, the use of the L1 norm means LASSO regression selects one of the predictors and drops the other (weights it as zero). This has been argued to provide a more interpretable estimators (Tibshirani 1996).

## Q16. When is Ridge regression favorable over Lasso regression?

**Answer:** You can quote ISLR’s authors Hastie, Tibshirani who asserted that, in presence of few variables with medium / large sized effect, use lasso regression. In presence of many variables with small / medium sized effect, use ridge regression.

Conceptually, we can say, lasso regression (L1) does both variable selection and parameter shrinkage, whereas Ridge regression only does parameter shrinkage and end up including all the coefficients in the model.

In presence of correlated variables, ridge regression might be the preferred choice. Also, ridge regression works best in situations where the least square estimates have higher variance. Therefore, it depends on our model objective.