Control Bias and Variance of Linear Models of small sample size using Regularisation

Under the guidance of Gourab Nath, Faculty, Data Science, Praxis Business School, Bengaluru.

Bhavya Vemavarapu

Post-Graduation in Data Science

Praxis Business School

[bhavyasaisree911@gmail.com](mailto:bhavyasaisree911@gmail.com)

*Abstract*—this paper helps to obtain an intuitive understanding of one of the most fundamental concepts in Machine Learning: The Bias-Variance trade-off using Regularization. Linear models have been a critical part of machine learning, more like a foundation, since the beginning of it. With the introduction of new models, many new problems came into existence, multicollinearity, bias-variance trade-off, sample size selection, and feature selection is some of the problems. The concept of bias-variance trade-off provides a mathematical basis for understanding the common modeling problem of UnderFitting vs. OverFitting. The bias-variance trade-off can be considered as a ‘sweet spot' in between an underfit model and an overfit model. This is something which is extremely difficult to achieve but not impossible. The motive of this paper is to pitch different ways in which bias-variance trade-off can be controlled. This experimentation of controlling bias-variance is done by exploring different scenarios which lead to overfitting. The results of this experiment will be used to present solutions from a different perspective for each of the problems mentioned. The linear model taken under consideration for experimenting is the linear regression model. This study focuses on the following problems; choice of the sample size for training the model, and the degree of a polynomial fit. Each problem is broadly taken under study and explored.

Key Words: Bias-Variance Trade-off, OverFitting, UnderFitting, multicollinearity, Regularisation, LASSO, RIDGE Regularisation.

**Introduction**

The main objective of this paper is to demonstrate the control over Bias and Variance in a linear model. In linear regression, the main objective is to minimize the cost function by selecting an appropriate value of ꞵ **curl.** Felicitous value ofꞵ **curl** reduces the error to minimum.

Y = b0 + b1X1 + b2X2 + εi

Minimum error of a model implicates that it is able to generalize the spread of data very well and make accurate prediction for forthcoming, unseen data. Linear model does not perform better with the factors such as high number of feature, sample size, multicollinearity and polynomial fit of higher order. The accuracy of a linear model decreases with increase in any of the former mentioned factors except sample size of a dataset. That is, model complexity of a linear model is directly proportional to the factors mentioned. The idea of linear regression is build a model on known dataset, train dataset, and try to make prediction on data which has not been seen by the model. This unseen dataset is usually referred to as test dataset. When a model is trained on train data, it means that the model can now generalize the train data very well and can make a good prediction of test data based on this existing data. But there are chances that if an unseen data, say test data, are shown to the model it cannot make good predictions, that is it would make lots of wrong prediction. These wrong predictions are made because the model has memorized the pattern of the train dataset which is does not resemble the pattern of the test data. This is called over-fitting. Figure (i) show the overfitting of a model on actual data set.

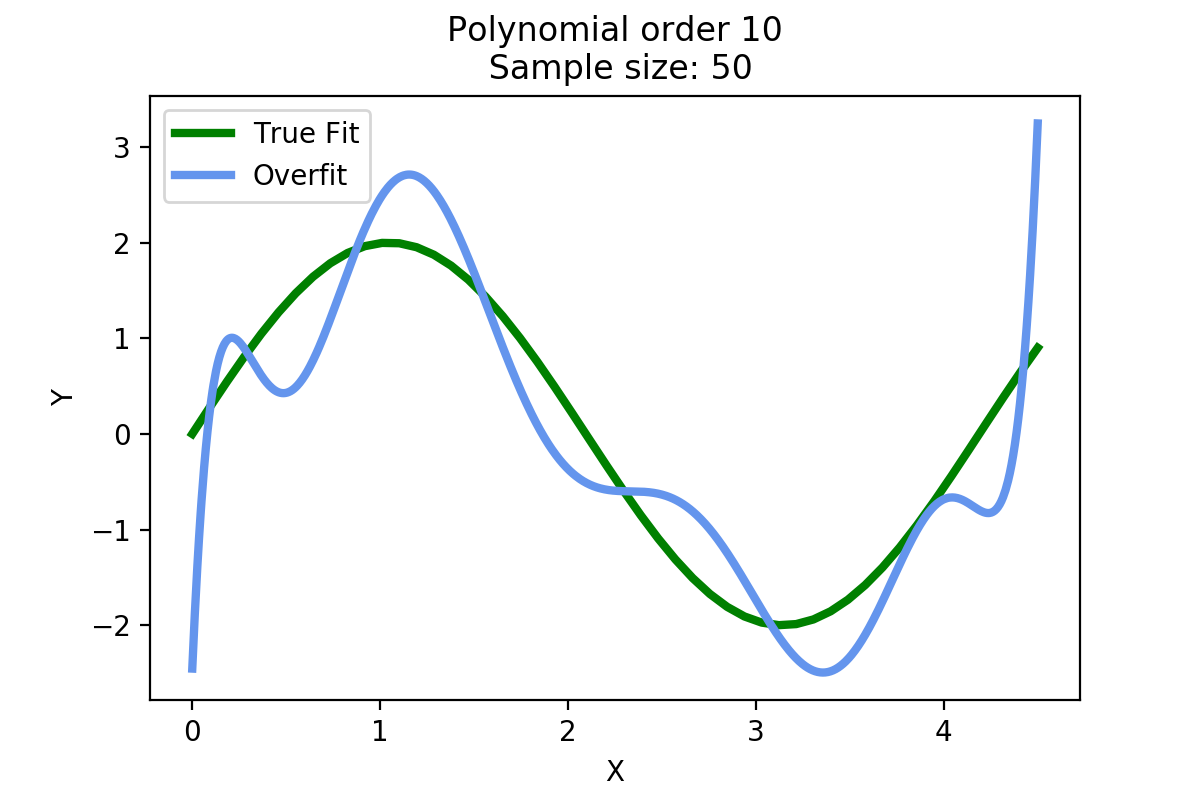


Figure (i): Linear regression

The vast difference in true fit and overfit is caused by some error. This error can be decomposed into three sources of error:

1. Irreducible Error
2. Bias
3. Variance

The Expected prediction error can be written as:

Error(x) = σ2 + +

An irreducible error is something that cannot be explained, it’s the noise in a data. Since it’s near to impossible to reduce this error we focus on reducing the bias and variance.

To understand bias and variance we first need to understand the reason for overfitting in models. We will be experimenting with a model trained on a small sample and a model with a polynomial fit. The dataset used for experiments is created by using the function:

Y = b0 + b1X1 + b2X2 +…………+b10X10+ εi

A sample is created and noise is manually added to it to have a better understanding of this concept. All the models are trained on the sample from the dataset containing noise.

What is the best polynomial fit?

Polynomial regression is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modeled as an nth degree polynomial in x.

Y = b0 + b1X1 + b2X2 +…………+bnXn + εi

Figure (ii) demonstrates polynomial fit up to order 6 with the sample size and sample data kept as constant.

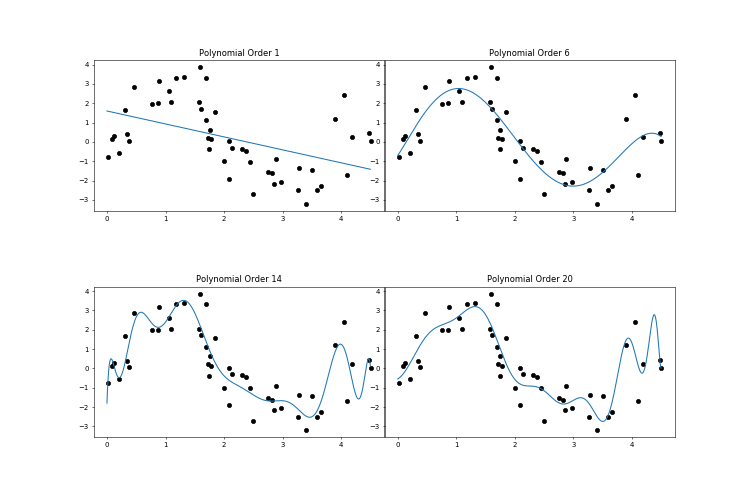


Figure (ii): Linear fit of different polynomial order

We can see that as the polynomial order increase, the complexity of the model also increases, i.e.) the model is getting better and better with every next polynomial order. The model is explaining and capturing almost all the data points. But the drawback to this is that this model might fail to explain any unseen dataset. Figure (iii) shows the polynomial fit of different on the actual dataset. They differ a lot from the true regression line.

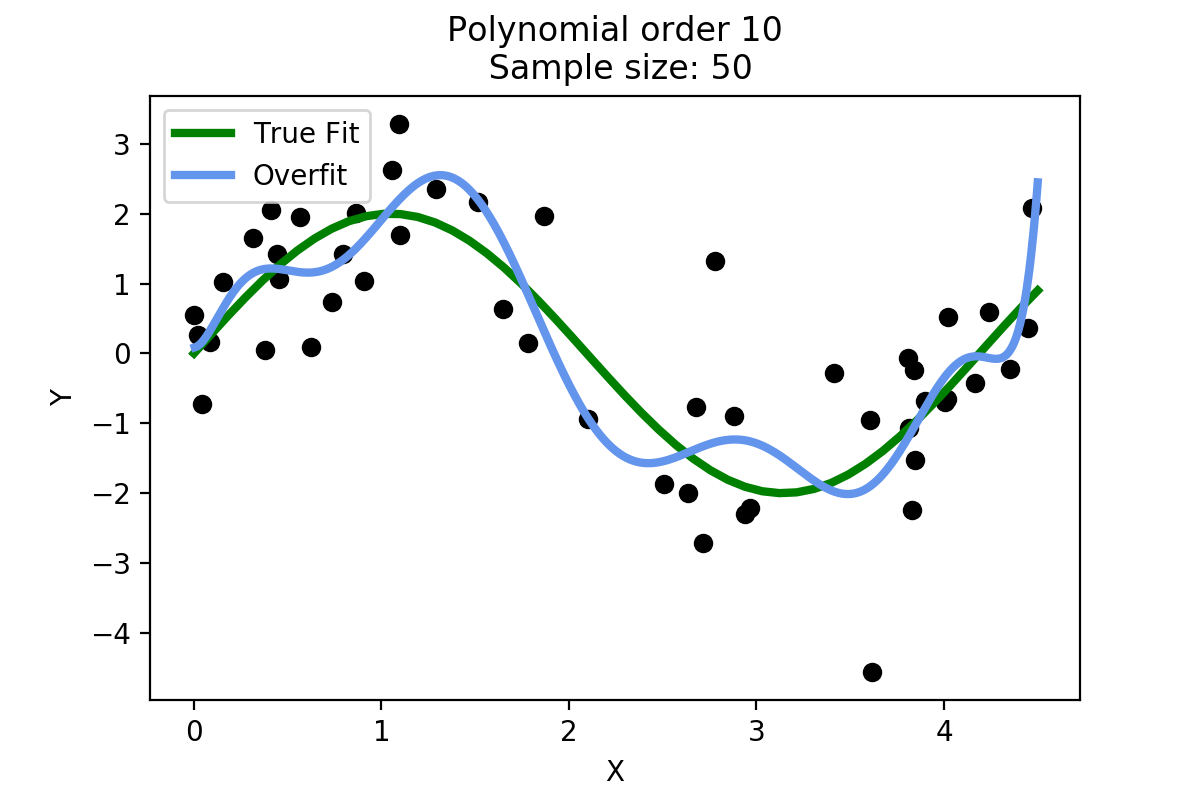


Figure (iii): True fit of data vs. polynomial fit order 10

Figure (iv) shows us the train error and test error plotted in a graph. As we move across the x we can see that train error decreases significantly. On the other hand, the train error increases by a lot. This means that as the model complexity increases, the model is able to generalize the train data set really well. The more a model generalizes the data, more it will fail to predict on an unseen dataset.

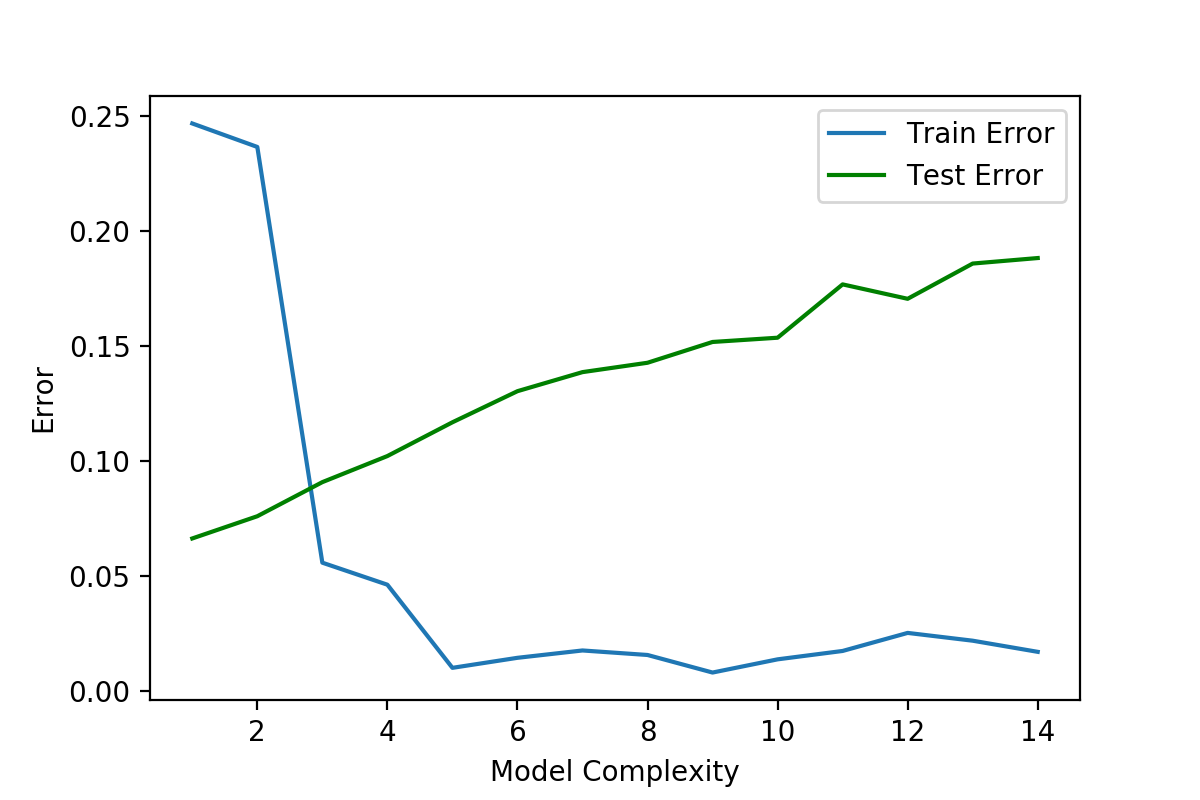


Figure (iv) Test error vs. Train error

B. Introduction of Bias and Variance

We have witnessed that with small sample size and higher order polynomial has high error rate of a model. In order to reduce this error rates we focus on reducing bias and variance because this values are quantifiable unlike irreducible error.

Bias: It quantifies how far is average of all models fitted over all possible existing data set from the true regression line. Consider ith observations (xi, yi)

Bias at xi~ = [f β~ (xi~) – Etrain [f β~^ (xi~)]]

Variance: It is a measure of spread or variations in our predictions and it also measures an estimate varies around the average fitted regression line.

Variance at xi~ =

Etrain [f β~^ (xi~) - Etrain [f β~^ (xi~)]] 2

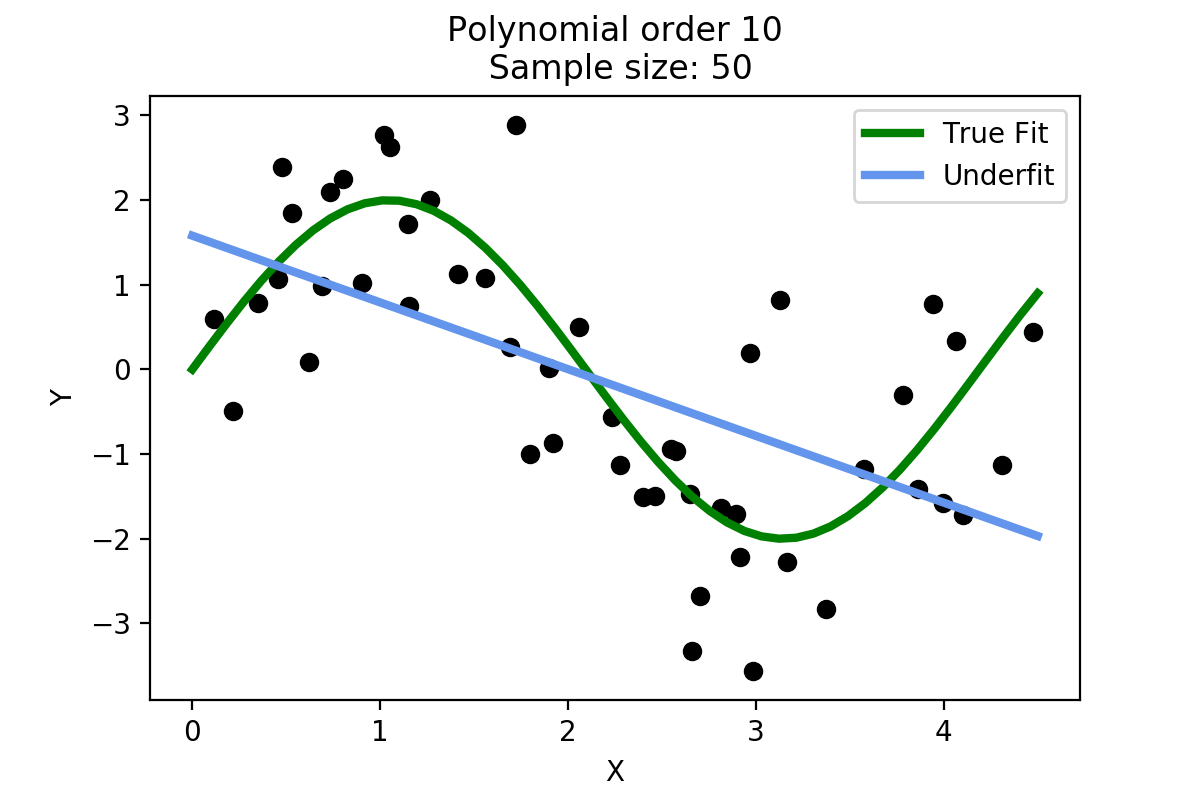
Figure (v) shows underfitting. Underfitting happens when a model unable to capture the underlying pattern of the data. These models usually have high bias and low variance. It happens when we have very less amount of data to build an accurate model or when we try to build a linear model with a nonlinear data. Also, these kind of models are very simple to capture the complex patterns in data like Linear and logistic regression.

Figure (v): Underfitting model

Figure (iii) shows overfitting, overfitting happens when our model captures the noise along with the underlying pattern in data. It happens when we train our model a lot over noisy dataset. These models have low bias and high variance. These models are very complex like Decision trees which are prone to overfitting.

If a model is too simple and has very few parameters then it may have high bias and low variance. On the other hand if our model has large number of parameters then it’s going to have high variance and low bias. Thus there is a need to find the right/good balance without overfitting and underfitting the data.

This tradeoff in complexity is why there is a tradeoff between bias and variance. An algorithm can’t be more complex and less complex at the same time.

In bias and variance trade off, a model gives up either on bias or variance to get the accurate model according to the requirements.

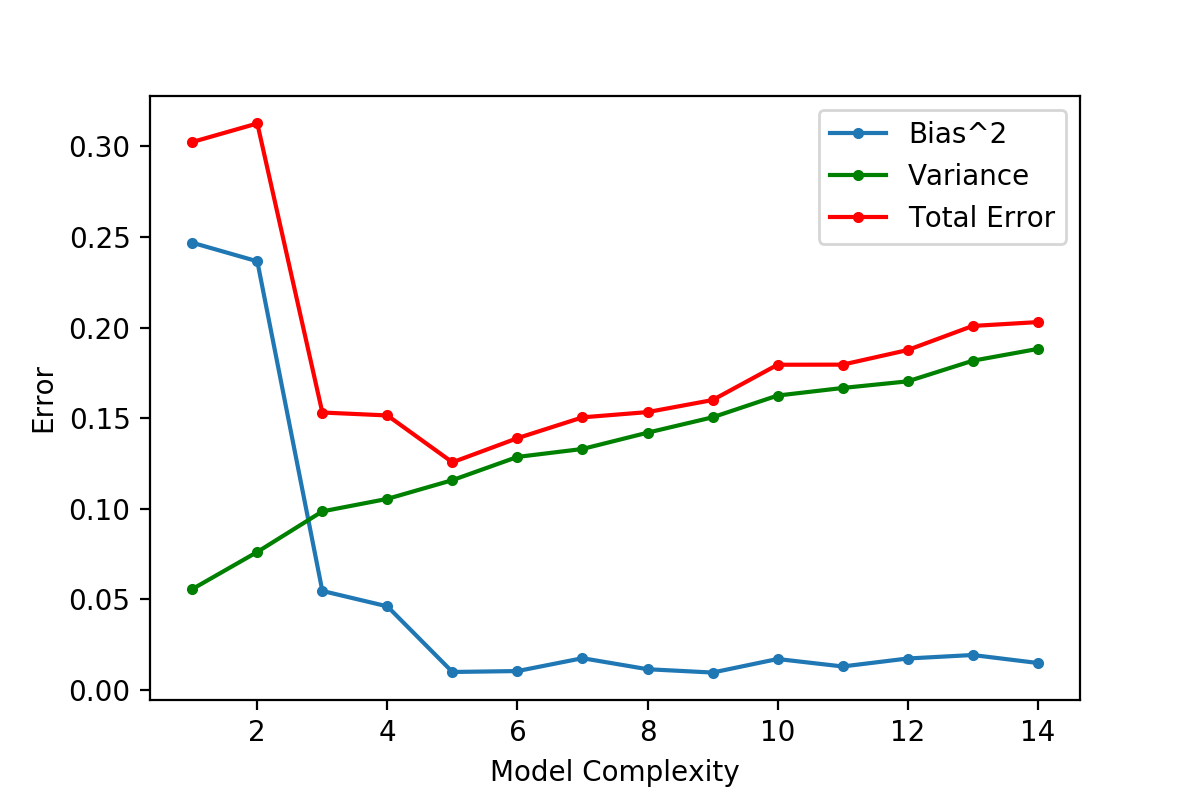


Figure (vi): Bias Variance trade-off

Figure (vi) show the bias variance trade off. It is the sweet spot just before the variance starts to increase and bias stars to drop. This spot is where the model has a good skill on both existing data and unseen data. We can notice that the model complexity of 3 would be the best out of all the other models as the total error of this model is minimum while minimizing the bias and maximizing the variance.

REGULARISATION

In linear regression, we choose β coefficients which minimizes the cost function and which reduces the error, this helps to fix the data well. But this doesn’t help in controlling the overfitting of the model. When the β coefficients are unconstrained, they can tend to have a high value which would result in high variance in the model. In order to control variance and bias, we add a penalty parameter lambda (λ) to the β coefficients while estimating them. This type of regression where we add penalty parameter in order to estimate the β coefficients is called Ridge and Lasso Regression.

Ridge Regression (L2 Regularisation): It helps us automatically reduce Bias and Variance. Ridge regression tries to avoid overfitting of data by penalizing large coefficients. It uses L2 regularization to impose a penalty.

Ridge Cost Function = - 2 +i2

λ – Tuning Parameter, this parameter controls the amount of shrinkage. The value of λ lies between 0 and ∞.

The modified ridge cost function which will help us to control two things:

1. Control the fit of the model.
2. Control the magnitude of coefficient estimates which control overfitting

Here λ-lambda helps us to tune the cost function i.e.

- 2

If λ is high, the overall cost function increases. In order to minimize this cost function, the optimization problem chooses a very small value of β. We will choose λ correspondingly to minimize the mean square error.

If λ is ∞, β coefficients become Zero and greater will be shrinkage of the β coefficients which will make the coefficients more robust to collinearity.

If λ is Zero, β coefficient estimates will be equal to normal Ordinary Least Square i.e.) linear regression solution.

The another way to choose the value of λ in order to control bias and variance is Ridge Regression Cross Validation is used to ensure that the best model is selected.

In Ridge Regression Cross Validation technique, the steps to follow:

1. Create a grid for the value of λ
2. Use cross-validation and loop that cross-validation technique to choose the value of λ which minimizes the error on new data.
3. Create a vector for new data error
4. Now create for loop, iterate the λ value and use cross-validation to find the new data error and append the error to the vector which created.
5. End the for loop
6. Now the model will give the best λ value which has the minimum predicted error.

The important note in Ridge Regression is L2 regression enforces β coefficients to converge to zero, as seen in figure (vii), but doesn’t make their value to zero. So ridge regression will not make the irrelevant variable coefficients to zero rather it will reduce the impact of these variables by putting small weights on them. Ridge regression allows us to keep all the variables by making adjustments.

Lasso Regression (L1 Regularisation): It helps us automatically select the useful feature and also does regularization. Lasso is the least absolute shrinkage and selection operator.

Lasso Cost Function = – 2 +

This modified cost function can help us to put larger weights on important variables and set the weight of the redundant

Variables to zero. In lasso regression, it enforces the β coefficients to zero.

The main difference between ridge regression and lasso regression is, in lasso it not only penalizes the higher β values it also forces the irrelevant variable coefficients to zero.

ILLUSTRATIVE EXAMPLE:

To demonstrate the working of Ridge and Lasso we take a sample of size 20 and build a model on linear regression and the perform regularization on it using Ridge and Lasso. Figure (vii) shows how a linear, ridge and lasso are fit over a sample of 20 data points.

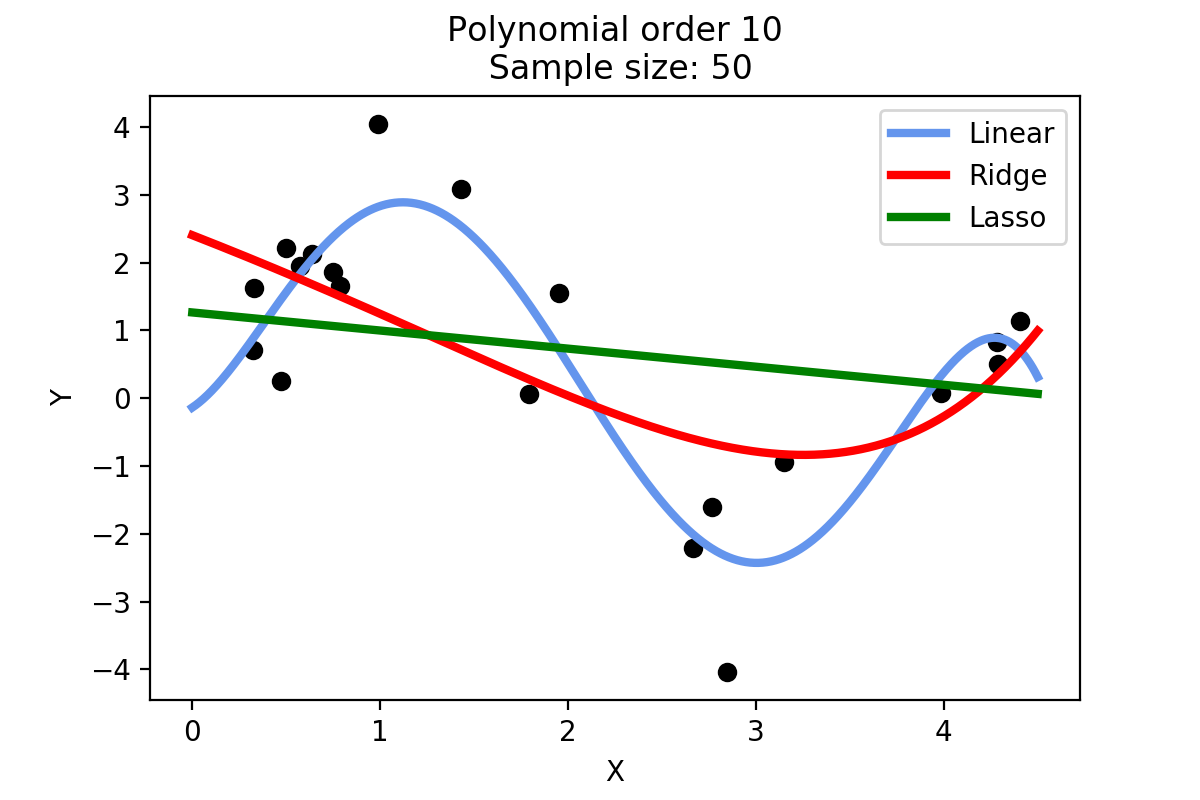


Figure (vii): Linear regression, Ridge and Lasso Regularisation.

CONCLUSION:

To sum it all, in order to get a better model we need to minimize the error. The cause for errors in model is due to overfitting or underfitting of the model due to inadequate sample or higher order polynomial fit. Error caused by such models is called Bias and variance respectively. To get the best model among several other models we need to select a value of bias and variance such that the model not only defines the spread of data really well but can also work the best on unseen data. In order to get the best values of bias and variance we regularize the linear model with Ridge and Lasso Regularisation.

References:

Professor Notes by Gourab Nath

Data mining and Predictive Analytics by Chantal D. Larose and Daniel T. Larose

<http://www.cs.columbia.edu/~blei/fogm/2015F/notes/regularized-regression.pdf>