* Ridge Regression
* Lasso
* Elastic Net
* Practical Examples

Linear regression which is popularly the OLS Regression is a method developed and used extensively during the late 1800s to early 1900s time frame. However, with advancements of computing technology, regression analysis can be calculated using a variety of different statistical techniques which has led to the development of new tools and methods.

In modern data analysis, we find often data with a very high number of independent variables and we need better regression techniques to handle this high dimensional modelling.

**Review of Linear Regression Analysis**: **Multiple Linear Regression**

Consider the following equation:

+ + E

A multiple linear regression is essentially the same as the simple linear regression, using the table below the interpretation can be thought of:

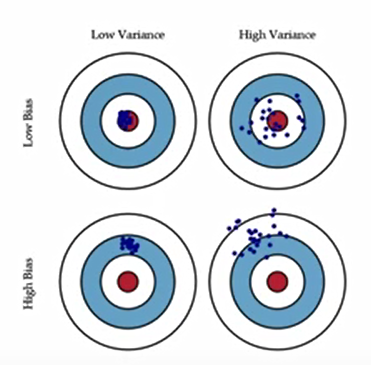
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Coefficients** | | | | |
| Term | Coefficient | Standard Error | T Value | P Value |
| Intercept | -114.326 | 17.4425 | -6.5544 | 0.02 |
| Height | 106.505 | 11.55 | 9.22 | 0.001 |
| Width | 94.56 | 8.345 | 5.6612 | 0.0048 |

For each 1 unit of change in width, increases Y by 94.56 ***but, this is while holding all other coefficients constant.***

**Ordinary Least Squares:** Try to minimize the distance between the data points (actual) and the corresponding point on the line.

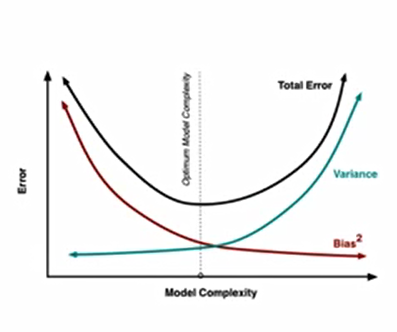
**Understanding the Error:** Errors in linear regression are essentially of two types – error due to bias and error due to variance. Though the understanding of error due to bias and error due to variance is important but in general what you really care is about the overall error and not the specific decomposition. This in turn helps to provide increased accuracy for our models.

**Bias and Variance Trade-offs**: Error due to bias is taken as the difference between the expected (or average) prediction of our model and the corrected value.



**Error due to variance** is taken as a variability of a model prediction for a given data point. Imagine you can repeat the entire model building process multiple times. The variance is how much the prediction for a given point vary between the different realizations of the model.

There is a tradeoff between a model’s ability to minimize bias and variance. Understanding these two types of error can help us diagnose model results and avoid the mistake of over-or-under-fitting.



Bias comes from selection, so if we do not do selection and include more and more parameters in the model, the model becomes complex. As the complexity of the model rises – variance becomes our primary concern while Bias steadily falls.

**Gauss Markov Theorem**: The Gauss Markov Theorem states that among all linear unbiased model estimates, the OLS has the smallest variance. This implies that our OLS estimates has the smallest mean squared error among the estimators with no bias.

This embarks to a very important question – ***“Can there be a biased estimator with a smaller mean squared error?”***

Shrinkage Estimators: We will replace our OLS estimates Bk with something slightly smaller –

B’k = \* Bk

What this tells is that if A is 0 we get our OLS estimates back but if A is very large the parameter estimate approaches a minimal value (0).

A is referred to as the shrinkage estimator (ridge constant). In principle with a right choice of A we can get an estimator with a better mean square error. The estimator is not unbiased but what we pay for bias we make up in the variance. To find the minimum A by balancing the two terms we get the following:

A=

Therefore, if all the coefficients are large with respect to their variances we set the A to be very small. (Because they induce less error anyways). On the other hand, if we have a number of small coefficients, we want to pull them closer to zero by setting A to be large.

The formula was further refined and a conclusion was drawn as below: We set all the coefficients to zero if we fail an F-test with a significant level as determined by the experiment (bias comes into picture). If we pass the F-test, then we shrink the coefficients by an amount determined by how strongly F-Statistic protests the null hypothesis.

**Regularization**: The concept of controlling over-fitting in a flexible and tune-able manner. Consider the following scenarios:

1. Linear Model – 2 data points – perfect line

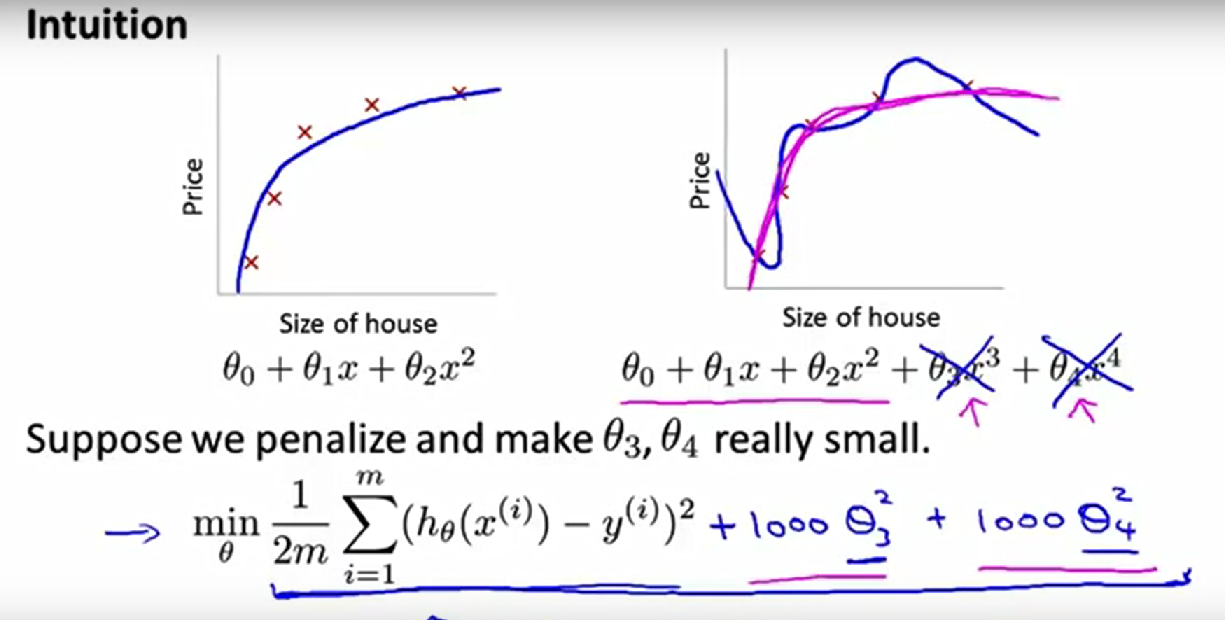
2. Quadratic model – two data points – infinitely many settings with zero control. How to choose?

3. High order coefficients = 0 – uses knowledge of where the features come from.

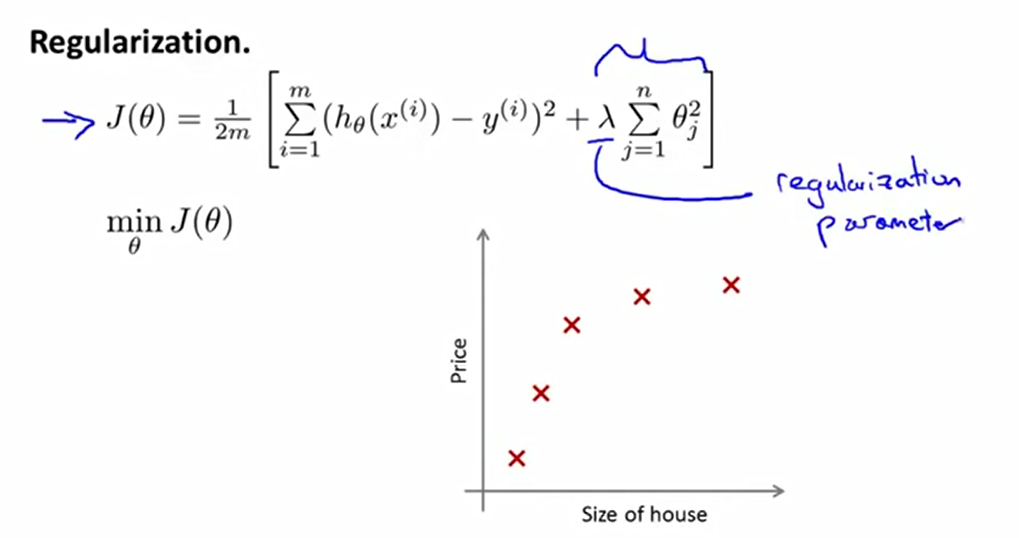
**What is over-fitting?** The problem of over-fitting happens if we have too many features, the learned hypothesis may fit the training set very well but would fail to generalize to new examples. (Predict prices on the new examples). Plotting the hypothesis might be the best way to address the problem of overfitting but plotting becomes difficult when we are addressing a problem of very large dataset with large number of features. The options to this are:

1. **Reduce the number of features**: either manually or through model selection algorithms.

**2**. **Regularization**: Keep all the features but reduce the magnitude or values of parameters – works well when we have a lot of features each of which contributes a bit to predicting y.

****

The idea of regularization is small value for parameters θ1, θ2, … , θn – simpler hypothesis with less prone to over-fitting.



**In this case the lambda is the tuning parameter. A very high value of lambda would result in very low or even close to zero values for the** θs. What this means is in effect we would have the equation equal to θ0 – classic case of under fitting in other words very high bias.

**Algorithms supporting Regularization**

**Ridge Regression**: Modelling technique that works to solve a multi-colinearity problem in OLS models through the incorporation of a shrinkage parameter.

The assumption of the Ridge Regression are the same as the OLS. Linearity, Constant Variance, and Independence. Normality need not be assumed. Additionally Multiple linear regression has no manner to identify a smaller subset of the important variables.

Now, one of the biggest obstacles of using the ridge regression is choosing the appropriate A. The inventors of the ridge regression suggested using a graphic which they called as Ridge trace. A ridge trace is a plot which shows the ridge regression coefficients as a function of A. ***When viewing a ridge trace, we are looking for A which the regression coefficients have stabilized.*** Often the coefficients vary widely for small values of A and then stabilize.