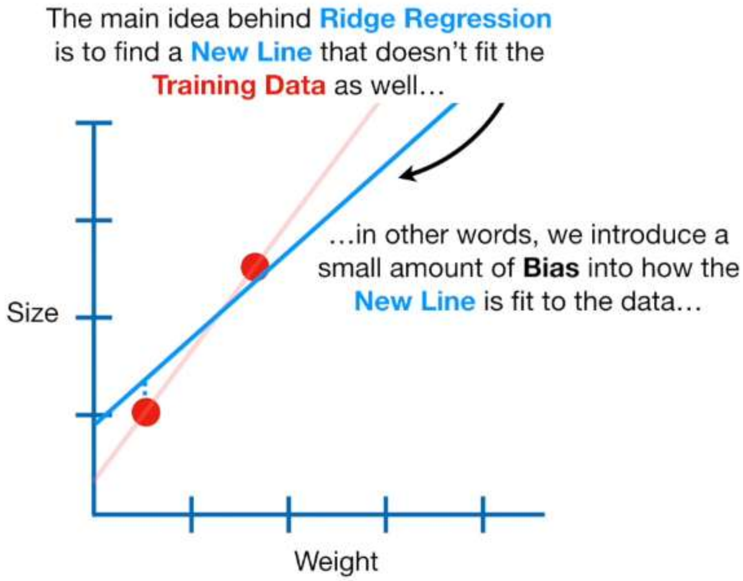
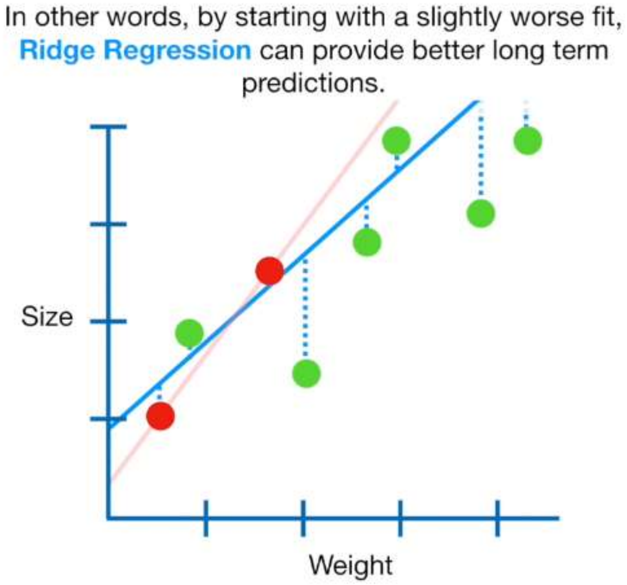
Ridge Regression

Ridge regression, also known as Tikhonov regularization, is a technique used in statistical regression analysis to deal with the problem of multicollinearity, where the independent variables are highly correlated with each other. It is an extension of ordinary least squares (OLS) regression.

In ridge regression, a penalty term is added to the OLS objective function to shrink the coefficient estimates towards zero. This penalty term is proportional to the square of the magnitudes of the coefficients, multiplied by a tuning parameter lambda (λ). The larger the value of λ, the greater the amount of shrinkage applied to the coefficients.

The objective function of ridge regression can be represented as:





1. Ridge in Simple linear Regression

It's used to reduce the slope of line or coefficient of line. In other words, we increase a small amount of Bias into how the new line is fit to the data.

Formula of simple linear regression $$\boldsymbol{y = \underbrace{m}_{\textit{Try to reduce } \boldsymbol{m\_slop}}x + c}$$

Lose function in simple ridge Regression: $\boldsymbol{L = \sum(y_i - \widehat{y_\iota})^2 \ + \ \lambda\, m^2}$

where:

- y is the Training variable,
- $\hat{y}$ is the Predicted variables,
- m is the vector of coefficients to be estimated,
- λ is the tuning parameter that controls the amount of shrinkage.

- Now derivate the lose function with respect to m and b

The Derivate with respected to b is: $\boldsymbol{b = \bar{y} - m\bar{x}}$

where:
- $\bar{y}$ is the mean of all y variable,
- $\bar{x}$ is the mean of all x variables,
- m is the vector of coefficients to be estimated,

Putting value of b in lose function:

$$\boldsymbol{L} = \sum (y_i - mx_i - \bar{y} - m\bar{x})^2 + \lambda m^2$$

$$\frac{\partial l}{\partial m} = \frac{\partial}{\partial m} (\sum (y_i - mx_i - \bar{y} - m\bar{x})^2 + \lambda m^2)$$

$$= -2\sum (y_i - mx_i - m\bar{x})(x_i - \bar{x}) + 2\lambda m$$

$$= \lambda m - \sum (y_i - \bar{y})(x_i - \bar{x}) - m(x_i - \bar{x})\hat{}2$$

$$m = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sum((x_i - \bar{x}))^2 + \lambda}$$

where:
- y is the Training dependent variable,
-x is the Training independent variable
- $\bar{y}$ is the mean of all y variable,
- $\bar{x}$ is the mean of all x variables,
- m is the vector of coefficients to be estimated,
- λ is the tuning parameter that controls the amount of shrinkage.

2. Multiple linear Regression

$$\boldsymbol{L = \sum(y_i - \widehat{y_\iota})^2}$$

Multiple ridge regression, also known as multivariate ridge regression, is an extension of ridge regression that allows for the analysis of multiple dependent variables simultaneously.

It is used when there are multiple response variables that are correlated with each other and with the independent variables.

In multiple ridge regression, the objective is to estimate the regression coefficients that minimize the sum of squared errors for all the response variables, while also incorporating the ridge penalty term. The objective function can be represented as:

In Multiple regression is lose function is :

$$L = (xw - y)^T \ (xw - y)$$

where:

- Y is the dependent variable,

- X is the matrix of independent variables,

- w is the vector of coefficients to be estimated,

- λ is the tuning parameter that controls the amount of shrinkage.

Lose function for Ridge Regression

$$L = (xw - y)^T \ (xw - y) + \lambda \, w^T w$$

$$L = (x^T w^T - y^T) \quad (xw - y) + \lambda \, w^T w$$

Arrows Multiplication

$$= w^T x^T x w - 2 w^T x^T y + y^t y + \lambda \, \mathrm{w}^{\mathrm{t}} w$$

Derivate the equation

$$w = \left(\boldsymbol{x}^{\boldsymbol{T}} \boldsymbol{x} + \boldsymbol{\lambda} \, \mathbf{I}\right)^{-\mathbf{1}} * \boldsymbol{x}^{\boldsymbol{T}} \boldsymbol{y}$$

where:

- Y is the dependent variable,

- X is the matrix of independent variables,

- w is the vector of coefficients to be estimated,

- I is the Identity Metrics

- λ is the tuning parameter that controls the amount of shrinkage.

3. Gradient Descent Ridge

Gradient descent ridge regression, also known as ridge regression with gradient descent, is a variant of ridge regression that utilizes the gradient descent optimization algorithm to estimate the regression coefficients. It combines the concept of ridge regression with the iterative nature of gradient descent to find the optimal values of the coefficients.

In gradient descent ridge regression, the objective is still to minimize the sum of squared errors, but with the additional ridge penalty term. The gradient descent algorithm is employed to iteratively update the coefficient estimates by taking steps in the direction of steepest descent of the objective function.

The steps involved in gradient descent ridge regression are as follows:

1. Initialize the coefficient values (β) to some arbitrary values.
2. Calculate the gradient of the objective function with respect to the coefficients.
3. Update the coefficient estimates by taking a step in the direction of the negative gradient, multiplied by a learning rate (α).
4. Repeat steps 2 and 3 until convergence or a predetermined number of iterations.

The ridge penalty term is typically incorporated into the gradient descent updates by adding the ridge penalty term to the gradient calculation. This penalty term helps to shrink the coefficient estimates towards zero and reduce the impact of multicollinearity. The learning rate (α) in gradient descent controls the size of the steps taken during each iteration. It is an important hyperparameter that needs to be carefully chosen. If the learning rate is too large, the algorithm may fail to converge, while if it is too small, the convergence may be slow.

Gradient descent ridge regression can be computationally efficient, especially when dealing with large datasets or a large number of predictors. However, it requires careful tuning of hyper parameters, such as the learning rate and the regularization parameter λ, to ensure convergence and find the optimal solution.

It's important to note that there are other optimization algorithms available for ridge regression, such as coordinate descent and singular value decomposition (SVD), which may offer advantages in different scenarios.

$$\boldsymbol{L = (xw - y\ )^T\ (xw - y)}$$

$$\boldsymbol{W_{new} = W_{old} - \eta \frac{\Delta \mathrm{L}}{\Delta w}}$$

$$L = \frac{1}{2}(x^T w^T - y^T\ ) \quad (xw - y) + \frac{1}{2}\lambda\, w^T w$$

Derivate the Equation:

$$\boldsymbol{\frac{\Delta L}{\Delta w} = \ \ x^T xw\ \ - x^T y + \ \lambda\, \mathbf{w}}$$