**Problem 6: Interpreting Results**

The factors to be considered while trying to decide the best method would be

* Nature of the data set itself,
* Accuracy
* Stability and
* Performance.

Nature of the data set refers to the volume and variety of the dataset and the number of features in it. Accuracy is the measure of correctness of the predictions made by the learning model. Stability and Performance is the behaviour of the model when new data points are introduced.

For this project, the diabetes dataset has only 242 training examples and 200 examples in test data which is not very large. So we consider Accuracy as an important factor to be taken into consideration when trying to compare the different methods. For this purpose, we can consider the RMSE values (Error Values) which is the mean squared error.

In the table below, the RMSE values have been calculated using optimal λ value.

|  |  |  |
| --- | --- | --- |
| **Method Used** | **RMSE - Training** | **RMSE - Testing** |
| **OLE** | 46.76708559 | 60.89203717 |
| **Ridge Regression** | 53.7038339189 | 56.2805719779 |
| **Ridge Regression using Gradient** | 53.646550611 | 56.4036097492 |
| **Non Linear Regression** | 62.80663049 | 62.32613124 |

On comparing the accuracies by considering the RMSE values for the above methods, **Ridge Regression** and **Ridge Regression using Gradient Descent** give minimal error for testing data than Non-linear and OLE methods.

But it could happen that the inverse of covariance matrix becomes unstable for large number of features, when we use derivative optimization technique. Also finding inverse of single covariance matrix is not possible. For these situations, **Gradient Descent Ridge Regression** method should be chosen over Ridge Regression method.