Linear Regression on the BIOSCAD data set

**Data sets:**

The pre-processed data (see pre-processing document) is split in to three variants. All three variants are based on the combined non-lesional data for both AD and control patients. This gave a total of 100 data points.

In the first variant, the continuous variables are logged before normalizing. In the second, the continuous variables are not logged before normalizing.

The third variant is based off the first (logged continuous data) but has a greatly cut down number of continuous attributes. This is because, in the original data set, many values are marked as being below the detection range. As such, they are not reliable values.

**Models:**

For data sets 1 and 2, elastic net regularization with varying values of both alpha and lambda was used. The data sets were split up to use 60% for training, 20% for cross-validation, and 20% for testing. The models were also trained on 100 different splits in the data to see how the best values of alpha and lambda varied.

For data set 3, generalized linear regression was used. The data set was split in to 80% training and 20% testing. Once again, 100 different splits of the data set were used to see how model performance varied.

Each model outlined above was created twice, once using objective SCORAD as the dependent variable, and once using total SCORAD.

**Performance:**

To calculate performance of a given model the root mean square error between the predicted result and the actual result was calculated. These are shown below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Logged  (Elastic net) | Unlogged  (Elastic net) | Reduced  (General linear regression) |
| Total SCORAD | 13.65 | 31.61 | 14.46 |
| Objective SCORAD | 10.78 | 15.75 | 10.77 |

We also calculated the number of ‘successful predictions’. This is based on an absolute difference of less than 9 points as this is the minimum clinically important difference. The percentage of successful predictions is shown below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Logged  (Elastic net) | Unlogged  (Elastic net) | Reduced  (General linear regression) |
| Total SCORAD | 52.7% | 42.5% | 36.9% |
| Objective SCORAD | 68.8% | 64.9% | 61.4% |

**Comparison to an average predictor:**

We also created a model that simply predicted the average value of either total or objective SCORAD. This model performed as shown below:

|  |  |  |
| --- | --- | --- |
|  | RMSE | Successful predictions |
| Total SCORAD | 15.31 | 25% |
| Objective SCORAD | 11.23 | 35% |

It is worth noting that if the MCID is increased to 10, then the percentage of successful predictions drastically increases to 83%. This is because all the control data samples (SCORAD of 0) will now fall within one MCID of the average.

Our best model, logged continuous data compared to objective SCORAD using elastic net, performs significantly better than the average predictor when compared by successful predictions. However, it is has a very similar RMSE score.