Linear Regression on the BIOSCAD data set

Introduction

**Overview:**

The aim of the project was to create a model that, based off the biomarkers in the skin, would predict the SCORAD of a patient. This was done based off the data from the AMC BIOSCAD study.

Method

**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. Data from lesional skin was not used as, after pre-processing, there were only 17 data points remaining.

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 only two of the continuous data attributes are used: IL-1a and IL1β. This is because, in the original data set, the remaining attributes have significant numbers of values marked as being below the detection range.

**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. Using the validation data, we chose the best alpha and lambda values. We then combined these using a weighted mean based on the performance against the testing data to find the overall best values.

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.

Results

**Alpha and Lambda parameter values:**

Alpha values close to 1 favour lasso regression, values close to 0 favour ridge. The best alpha values for each model are shown below:

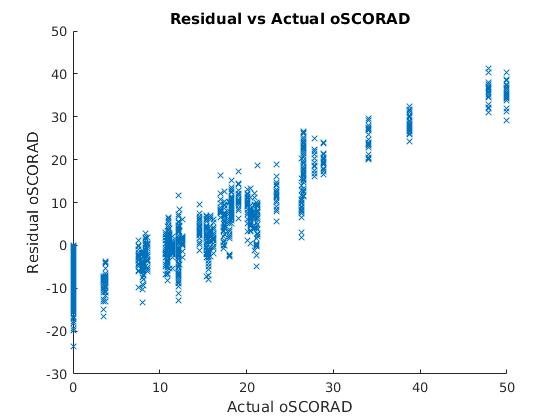
|  |  |  |
| --- | --- | --- |
|  | Logged  (Elastic net) | Unlogged  (Elastic net) |
| Total SCORAD | 0.376 | 0.624 |
| Objective SCORAD | 0.393 | 0.606 |

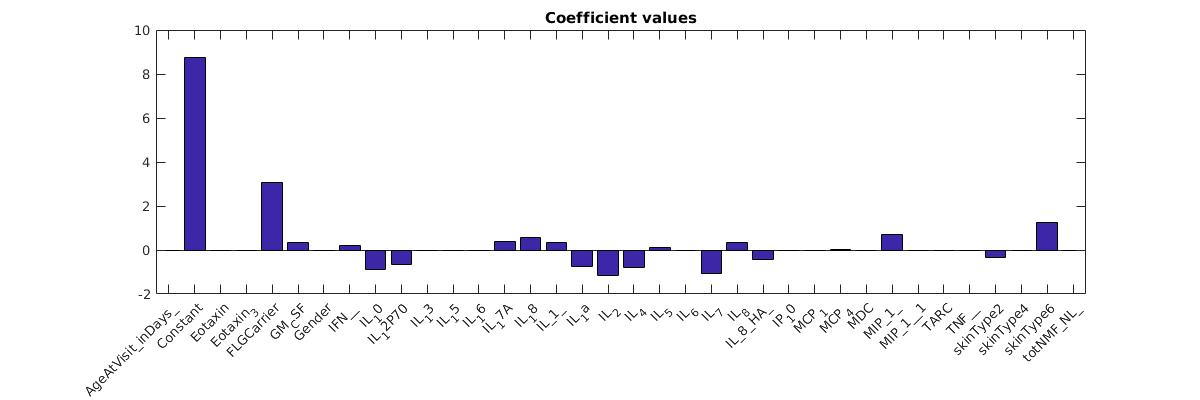
The best lambda values for each model are shown below:

|  |  |  |
| --- | --- | --- |
|  | Logged  (Elastic net) | Unlogged  (Elastic net) |
| Total SCORAD | 1.62 | 2.34 |
| Objective SCORAD | 1.73 | 2.12 |

**oSCORAD logged data results:**

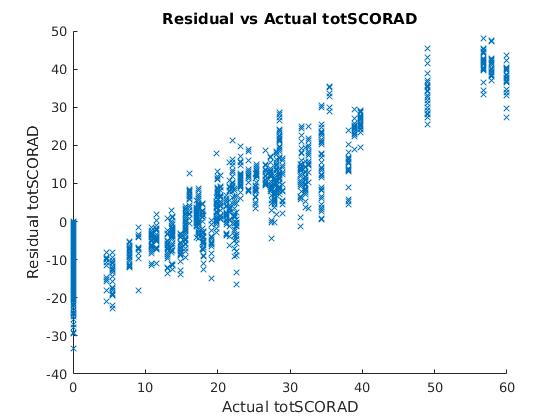


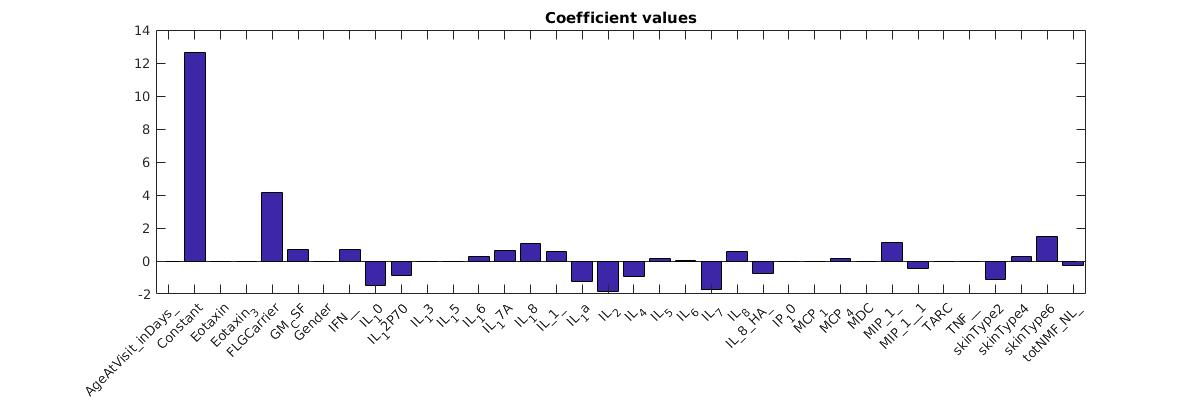




**totSCORAD logged data results:**

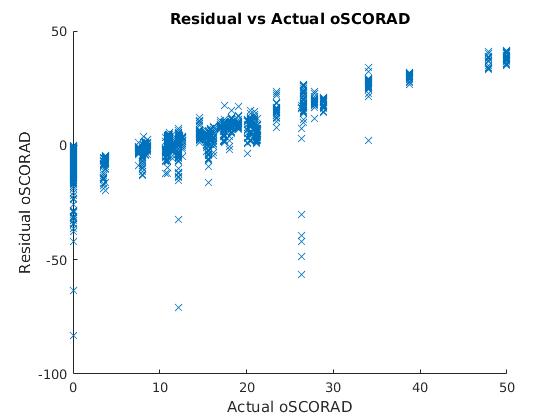


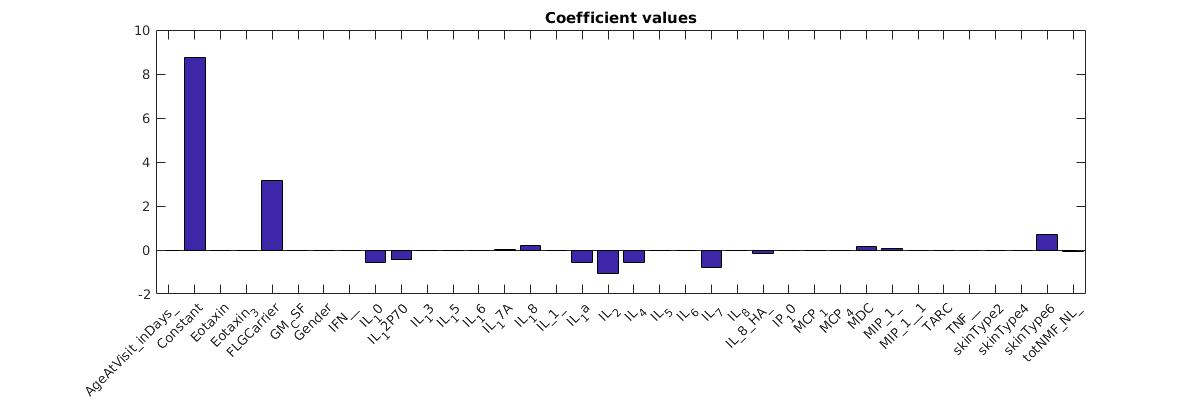




**oSCORAD unlogged data results:**

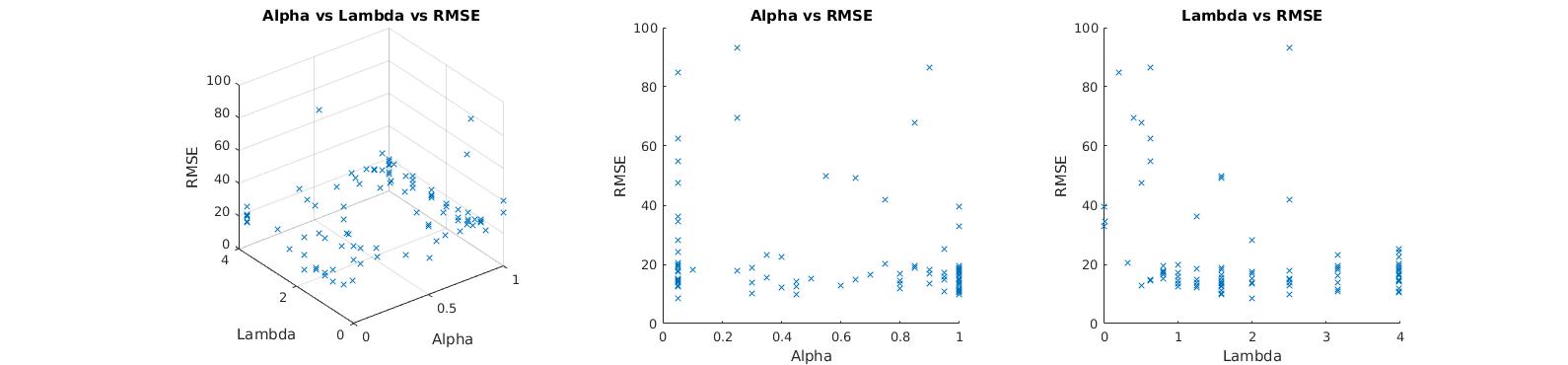
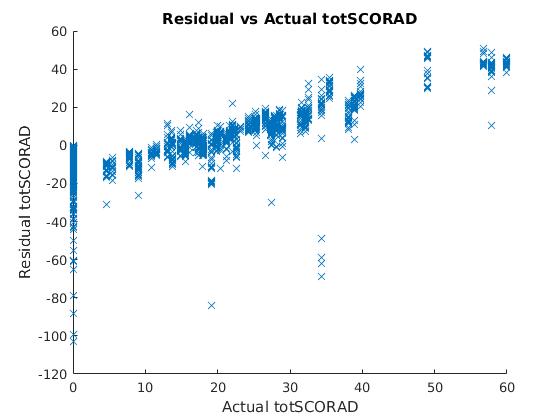


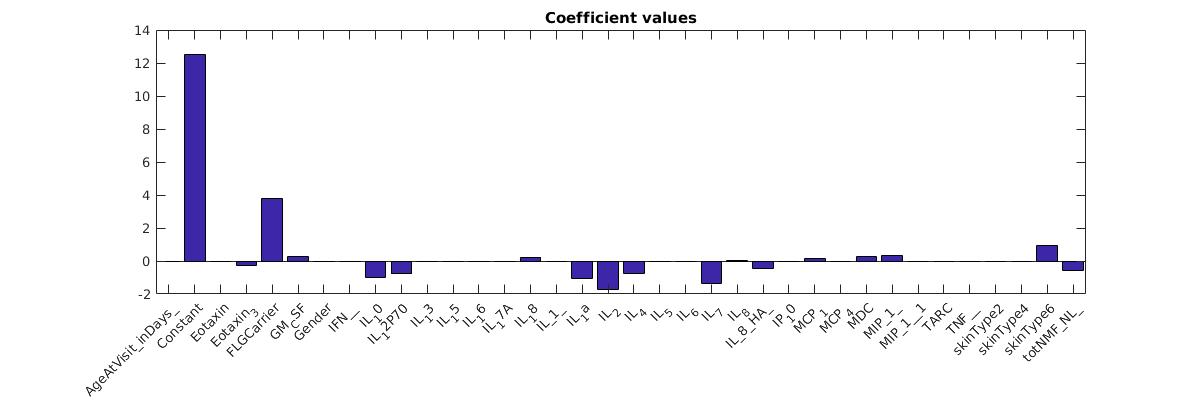




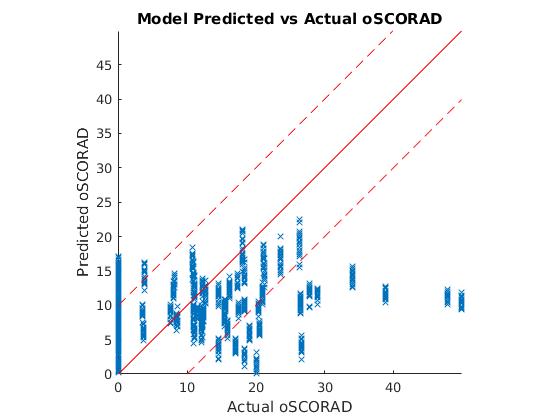
**totSCORAD unlogged data results:**



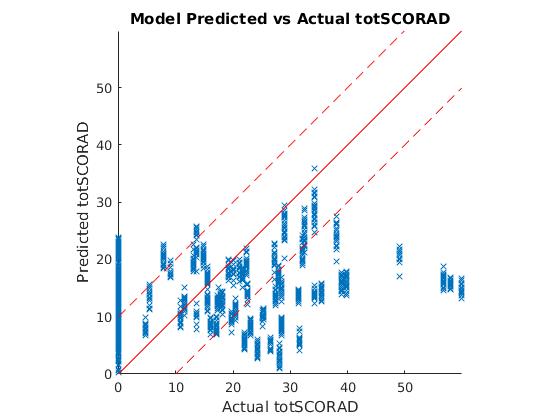




**oSCORAD reduced subset results:**



**totSCORAD reduced subset results:**



Discussion

**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 | 27.60 | 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 for objective SCORAD and 13 points for total SCORAD. These values were picked as they are 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% | 49.0% | 36.9% |
| Objective SCORAD | 68.8% | 64.9% | 61.4% |

As shown above, the models trained on logged data performed better than those based on unlogged data. This is likely because logging the data reduces the effect that large value outliers have. This is particularly important as the training data is small, thereby making it harder to spot outliers. Evidence of this can be seen in the very large predictions made in the unlogged models – likely a result of outliers in the training data.

In addition, models using objective SCORAD as the dependent variable performed better than those using total SCORAD. This is because it is very difficult for the subjective component in total SCORAD to accurately represent the severity of the disease as it depends on how the person feels. As such, the data that the model is trained on and compared to is less reliable.

**Coefficients:**

From the graphs, we see that in all models the largest coefficient by far is the constant. This shows that the model is predicting very similar values regardless of the input data.

**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 | 15.06 | 35.8% |
| Objective SCORAD | 10.76 | 58.8% |

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.

In order to compare the predictions to the null model, I computed the kappa coefficient relative to average predictors for null and total SCORAD accordingly. This is shown below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Logged  (Elastic net) | Unlogged  (Elastic net) | Reduced  (General linear regression) |
| Total SCORAD | 0.263 | 0.206 | 0.017 |
| Objective SCORAD | 0.243 | 0.148 | 0.063 |

As seen in the table, none of the models perform significantly better than the average prediction model. This is to be expected as the coefficients showed that the values predicted by the model differ only slightly depending on the inputs.

**GIGO:**

It is highly likely that the reason that the models did not perform well was due to GIGO (garbage in, garbage out). That is to say, the data used to train the models was poor. This is because, as mentioned earlier, in all but two of the continuous attributes the majority of the data was below the detection range. As such, it’s value is unlikely to be accurate. As the model is being trained on inaccurate data, it’s output values will be at least as inaccurate.

In addition, the data set was too small. Of the 100 data points, only 53 were from patients with atopic dermatitis. Once testing and validation sets were extracted, only 60 points remained to train the model. Given that there are 35 attributes, this is simply not enough data to be able to train a model.