Analyzing the relationship between fMRI and Alzheimer's Progression via Machine Learning

Eric Boivie
Jimmy Boivie
Nicholas Cavanna
Liam Dickinson
Sam Garfinkel
d Connor Riley

May 5, 2015

1 Introduction

For this project, we were to investigate the relationship between Alzheimer's patients' functional magnetic resonance imaging (fMRI) images and their scores on the Mini Mental State Examination (MMSE). Furthermore, we were tasked with training a regression model on a training sample of data of fMRI images and the patients' respective MMSE scores via multiple machine learning methods, and then using the regression model to estimate MMSE scores for a test set of data. Specifically, the training sample size was 800, each representative of an Alzheimer's patient with 285 $\mathcal{N}(0, 1)$-normalized features representing fMRI image data along with a MMSE score. The test sample consists of 200 data sets of features and a corresponding MMSE score as well.

We chose to train four different regression learning models, namely Support Vector Regression (SVR), Bayesian ridge regression, Lasso linear regression, and ridge regression. We used the Python programming language module SciKit-Learn to implement each of these regression models. In order to tune the parameters for these regression models, we used a grid search as provided by SciKit-Learn in order to search through a parameter space. We utilized the cross-validation technique Generalized Cross-Validation, a form of efficient Leave-One-Out cross-validation. We changed the parameter space until we narrowed in on optimal parameter(s) for each model.

2 Analysis Strategy

In order to analyze each of the models' performance in an easy comparative manner to determine which performed the best for this scenario, we first computed the root-mean-square error, which is

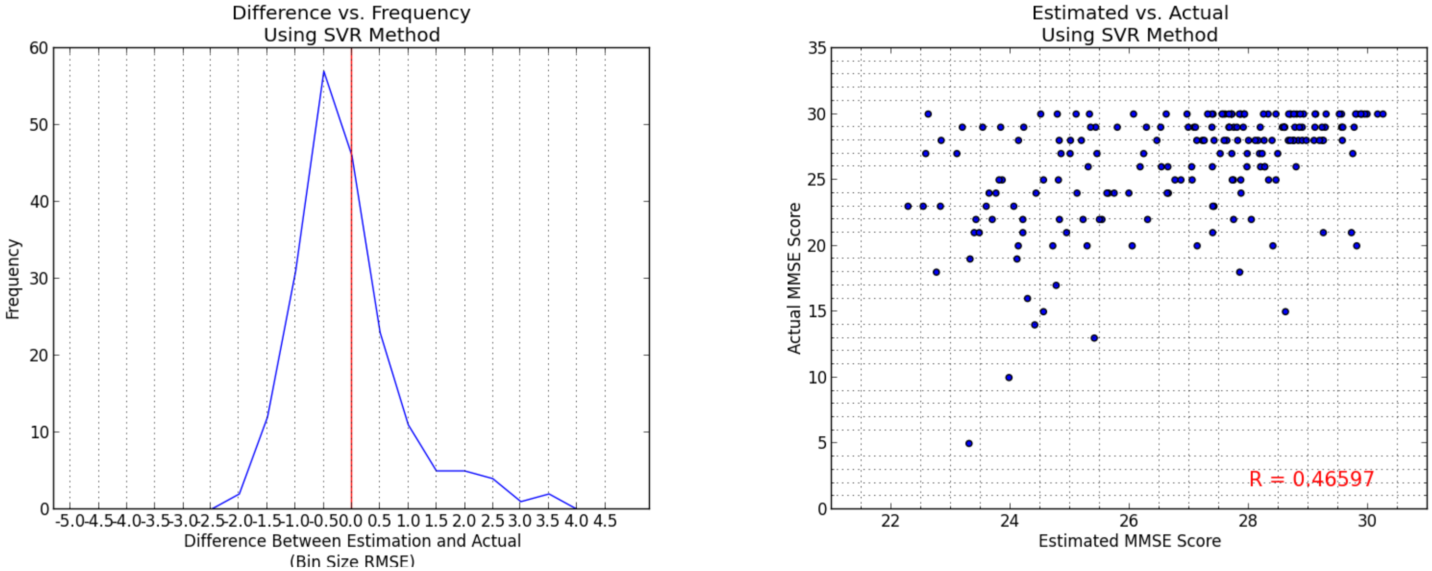

Figure 1: Distribution differences between estimated and actual MMSE scores for SVR

calculated as follows.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{x_i} - x_i)^2}{n}}$$

Each $\hat{x_i}$ represents the expected MMSE score for the testing data based on the regression model, and x_i is the corresponding actual score for the patient i in the testing data. This is a generalized standard deviation used to measure the spread of the results in a regression. We then took the differences $x_i - \hat{x_i}$ for each of the 200 patients in the testing set and binned them in $.5RMSE$-length intervals in order to create a histogram that would represent the distribution of differences between the estimated set based on the regression model and the actual MMSE values of the test set. The histograms are created so that each frequency per difference marker x counts all occurrences in the interval $[x, x + .5)$ i.e. the counts are left justified.

For the second part of the analysis, we plotted the calculated/expected MMSE scores for the testing data calculated via the regression model against the actual MMSE scores of the participants in order to calculate the correlation coefficients for each model.

We constructed both graphs from the data for each of the four methods using Python code in conduction with the imported library MatPlotLib.

3 SVR Results

Support vector regression is a modification of the classification method Support Vector Machine for regression problems. It separates the training data into two subsets via a hyperplane, one of

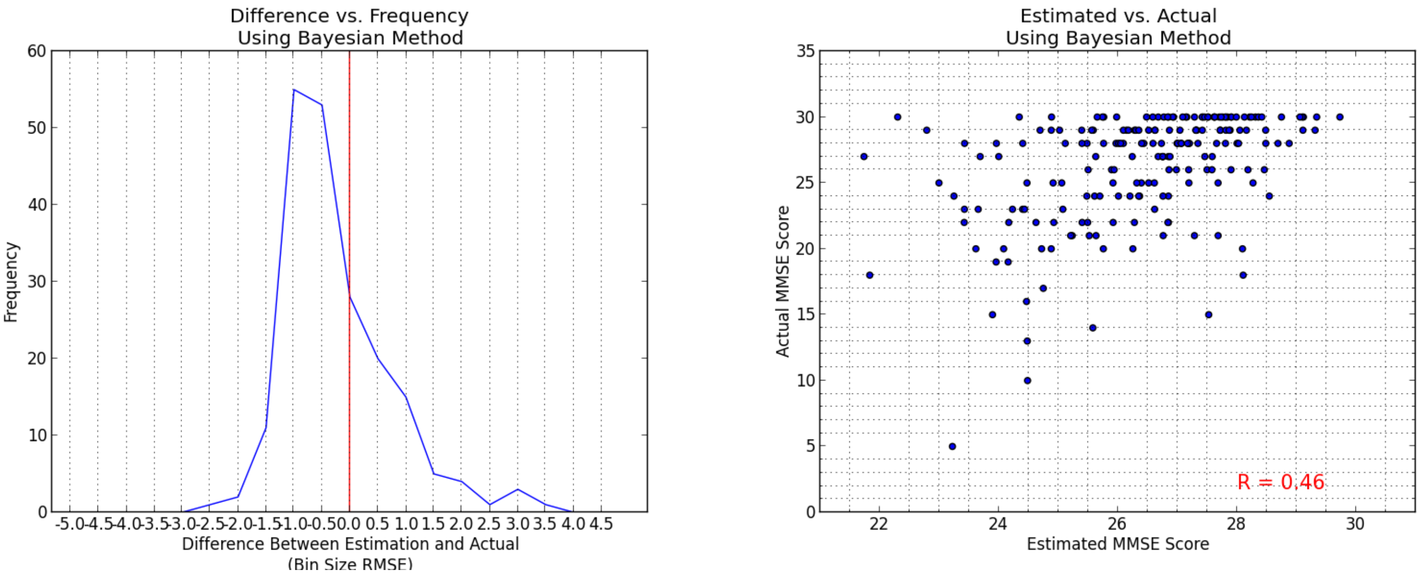

Figure 2: Distribution differences between estimated and actual MMSE scores for Bayesian

which is ignored in the model, as they are deemed to be unimportant in quantifying the relationship between the dependent and independent data in the training set.

The parameter C we trained using grid search that was used to create the regression model was $C = 3.45$.

The RMSE for the SVR model was 3.7805254167, with 78.5% of the testing data falling within 1 RMSE of the estimated, and 93.5% falling within 2 RMSE's.

Figure 1 depicts graphs of both our analysis techniques. SVR provided a slight underestimate of the MMSE scores of the testing data on average, based upon the histogram. The calculated correlation coefficient r based upon the scatterplot was .46597.

4 Bayesian Ridge Regression Results

Bayesian ridge regression utilizes Gaussian probabilistic models to fit the parameters $\alpha_1, \alpha_2, \lambda_1$, and λ_2 simultaneously based on the training data, so that it ideally fits well to the data, at the expense of computation time.

The parameters described above that we fit to the training set were $\alpha_1 = -51.5$, $\alpha_2 = -1 * 10^{-10}$, $\lambda_1 = -5$ and $\lambda_2 = -.03$.

The RMSE for the Bayesian model was 3.75664684144, with 78.0% of the testing data falling within 1 RMSE of the estimated, and 94.5% falling within 2 RMSE's.

Figure 2 depicts both the histogram and scatterplot graphs. Again, the Bayesian method slightly underestimated the MMSE scores of the testing data on average, and the calculated correlation coefficient was $r = .46$.

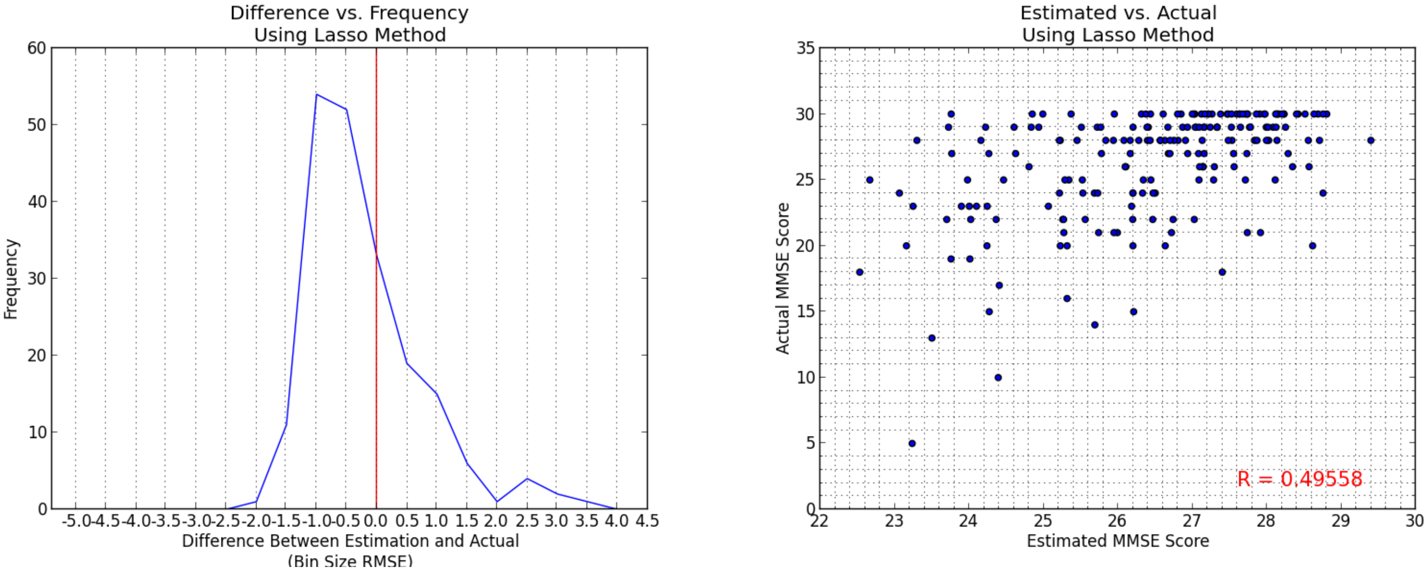

Figure 3: Distribution differences between estimated and actual MMSE scores for Bayesian

5 Lasso Linear Regression Results

Lasso is a linear regression model that minimizes the number of features on which the fit model is dependent by choosing those that display the most influence on the relationship between, in this case, the fMRI image values and the MMSE values for the training data. It does so by using a linear objective function and the l_1-norm on the parameter vector. Formally the function is as follows, where w is the parameter vector and α is the constant we find by fitting the regression model to the training set.

$$\min_{w} \frac{1}{2n} \|Xw - y\|_2^2 + \alpha \|w\|_1$$

The α constant we found using grid search and cross-validation was $\alpha = .1868$.

The RMSE for the Lasso model was 3.70743909872, with 79.0% of the testing data falling within 1 RMSE of the estimated, and 95.5% falling within 2 RMSE's. Figure 3 depicts both the histogram and scatterplots corresponding to Lasso.

The distribution of the difference between estimated and actual MMSE scores for the testing data were very similar to those obtained via the Bayesian regression method. The correlation coefficient for the Lasso was $r = .49558$.

6 Ridge Regression Results

Ridge Regression is similar in premise to Lasso regression, as the objective function for the coefficients is the same as Lasso's barring the use of the square of the l_2 norm on the parameter vector

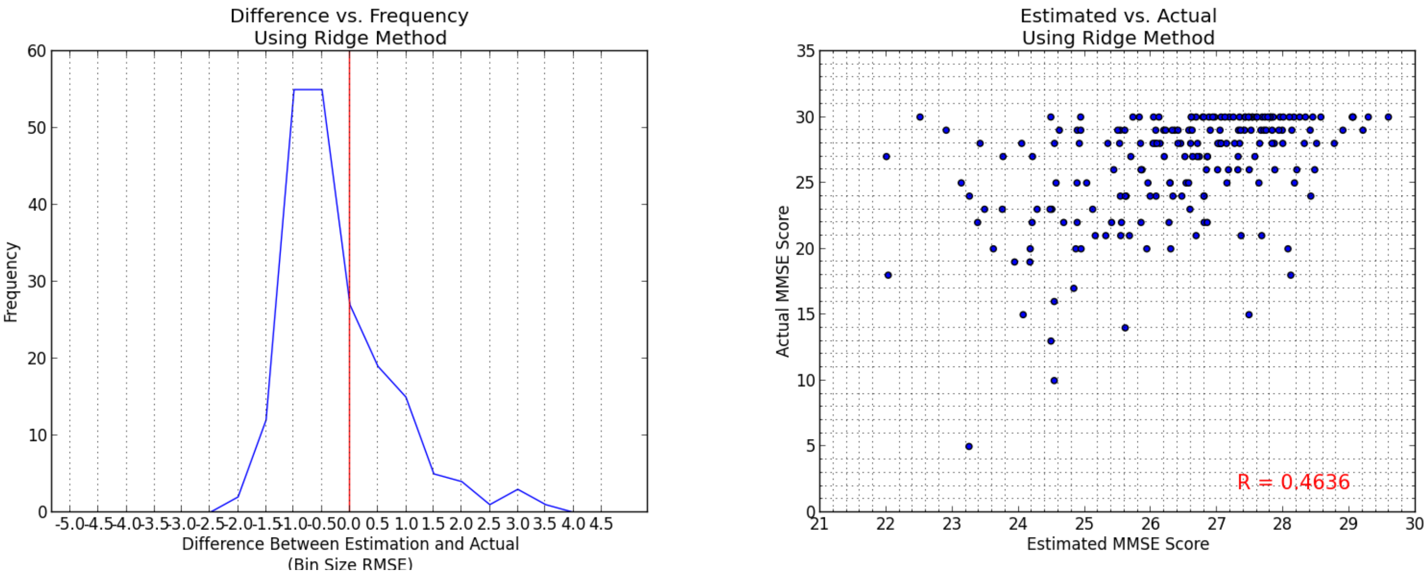

Figure 4: Distribution differences between estimated and actual MMSE scores for Bayesian

w and one does not divide by twice the sample size. It is as follows.

$$\min_{w} \|Xw - y\|_2^2 + \alpha\|w\|_2^2$$

α is again the parameter we had to fit to the training data. Ridge regression also has the same complexity as ordinary least squares, while making improvements to the optimization function.

The α constant we found during the fitting process was $\alpha = 1554.116$.

The RMSE for the ridge regression model was 3.75558117045 , with 78.0% of the testing data falling within 1 RMSE of the estimated, and 95.0% falling within 2 RMSE's. Figure 4 depicts both the histogram and scatterplots corresponding to Ridge regression.

Though the distribution in the histogram is favorable to that of the Lasso method, as it is more centered around 0, the correlation coefficient for the MMSE scores on the testing data was worse at $r = .4636$.

7 Conclusion

To conclude, we can see that the Lasso regression model managed to estimate the MMSE scores for the testing data based on the training set the best, with the correlation coefficient being .02961 higher than the next best, SVR regression. Assuming that we fit and ran the regression models correctly, it appears that since all of the models achieved approximately close correlation coefficients, and they were universally below .5, that fMRI image scores alone are an unreliable measure of the progression of Alzheimer's, or at least not over all stages of the disease. This is under the assumption that MMSE scores of patients correspond directly with their Alzheimer's progression, which based on its prevalence, is a reliable assumption. Though, the inaccuracy of these models in predicting the MMSE scores for the testing set could also be a result of the exact specifications of the fMRI data given, and a result of sample size and some of the exceptionally low scores in the testing data set.