Methods

Our data includes cortical, subcortical, and DTI measurements. All data was scaled using standard deviations from the mean. Analysis started with single variable regression against age, where the strongest correlations were linear and had a R^2 value just below 0.2. The five variables most highly correlated with age include superiorfrontal (R^2 = 0.192), GCC (R^2 = 0.169), parsopercularis (R^2 = 0.163), medialorbitofrontal (R^2 = 0.157), and thalamus (R^2 = 0.151).

Next, we determine which model for predicting age performed best. Models were tested with different input features selected. We see a linear distribution of ages among the training dataset, so cross validation with random sampling from a shuffled dataset is used for model assessment. 5-fold cross validation is used, giving training sets of 61 and test sets of 15. This cross validation was done 100 times on each model, with seeded random splits such that each model saw the same splits during training and fitting. For each cross validation, an average score was taken from the 5 configurations. Then, those averages were averaged after 100 iterations.

When using the full feature vector of 66 features, a Bayesian Ridge model performed best, averaging R^2=0.609 over 100 iterations of 5-fold cross validation, showing the best ability to generalize. When fitting the model to the entirety of the dataset, we see R^2=0.883. Other notable models using the full feature vector include SVM with RBF kernel (R^2=0.536), ElasticNet (R^2=0.593), and Ridge (R^2=0.516).

Subsets of the feature vector were also experimented with. These subsets include the cortical feature vector, subcortical feature vector, DTI feature vector, the top 5 variables linearly correlated with brain age, and the top 20 variables linearly correlated with age. The same models were trained in the same fashion described above, but no other subset of data performed as well as the full feature vector with a Bayesian Ridge Regressor model.

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **R^2 Linear** | **R^2 deg=2** | **R^2 deg=3** |
| superiorfrontal | 0.191806974 | 0.190957297 | 0.153828254 |
| GCC | 0.168573645 | 0.185481251 | 0.143588608 |
| parsopercularis | 0.162543893 | 0.148481573 | 0.107115252 |
| medialorbitofrontal | 0.156666067 | 0.144221257 | -0.102068301 |
| Thalamus | 0.151453842 | 0.154285246 | 0.126686093 |
| superiortemporal | 0.138638063 | 0.12829608 | 0.112861574 |
| rostralanteriorcingulate | 0.122889494 | 0.104477196 | 0.062824869 |
| CC | 0.109873238 | 0.11081502 | 0.085506722 |
| BCC | 0.105178546 | 0.099159928 | 0.059506878 |
| Left-Accumbens-area | 0.101227941 | 0.100850562 | -0.441981323 |
| FX | 0.084414875 | 0.075922113 | 0.042183436 |
| caudalmiddlefrontal | 0.078756231 | 0.033309392 | -0.097988119 |
| insula | 0.071186958 | 0.070307634 | 0.058560935 |
| supramarginal | 0.067991748 | 0.051994474 | -0.055696737 |
| frontalpole | 0.062361593 | 0.038629688 | -0.01095099 |
| rostralmiddlefrontal | 0.060488917 | 0.04040489 | -0.046457459 |
| parstriangularis | 0.059285932 | 0.062792963 | 0.027704727 |
| bankssts | 0.052170539 | 0.058758916 | -0.028086754 |
| CR | 0.022410172 | 0.01812194 | 0.010100561 |
| lateralorbitofrontal | 0.015881764 | -0.037889873 | -0.092436024 |

Figure: Table of the top 20 variables linearly related with age

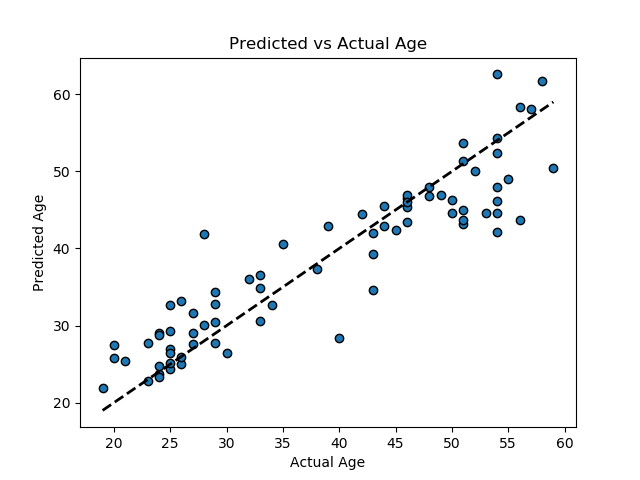


Figure: Predicted vs Actual Age for a Bayesian Ridge Regression model trained on all 76 samples (R=0.883)