Evaluating Polygenic Risk Score and Integrated Risk Model Methodologies for Alzheimer's Disease

Introduction

A polygenic risk score (PRS) is used to generate an individual's propensity to develop a disease given their genes. Through conducting a genome-wide association study, researchers can determine which parts of the genome, specifically which single nucleotide polymorphisms (SNPs), are significant for disease expression. A polygenic risk score is mathematically calculated using the weighted sum of the total amount of risk alleles an individual has for a disease. A polygenic risk score is often combined with other data about an individual, such as their age or sex, in order to provide the most accurate prediction for their disease likelihood, resulting in an integrated risk model.

The genetic components of Alzheimer's disease have been widely studied, as evidenced by the discovery of the APOE gene's role in Alzheimer's disease development. Polygenic risk scores are currently being generated for Alzheimer's disease, and there are various methodologies being used to do so. There still exist gaps in research in terms of determining which methodology generates a PRS or integrated risk model that is most predictive of an individual's likelihood of developing Alzheimer's disease, as well as determining ways to improve thesee predictive models. Through this project, I will evaluate various methodologies and assess which should be used to generate the best PRS and integrated risk model for predicting Alzheimer's disease.

The dataset used specifically in this project came from a plink file with data on 12,949 individuals obtained from the National Institute on Aging Genetics of Alzheimer's Disease (NIAGADS) database. The NIAGADS database contains a vast number of datasets with information on individuals both with and without late-onset Alzheimer's disease. Plink files contain crucial information about the phenotypes and genotypes of individuals, which is crucial for generating a polygenic risk score and creating an integrated risk model.

Methods

After using data preprocessing technqiues to read in the plink file and address missingness in the data, I generated the preprocessed dataset that will be analyzed. The preprocessed dataset contains 11,918 individuals, and it holds the following information on each subject: Family ID (FID), Sample ID (IID), affection status, the first ten principal component values, age, sex, APOE gene information, and ethnicity. Affection status corresponds to whether or not an individual has developed Alzheimer's disease, taking on a value of 1 or 0. Principal components, which were generated using principal component analysis (PCA), capture and explain the variability in the data, reducing down our high-dimension dataset to allow for greater interpretability. With APOE gene information, the dataset contains the number of E4 alleles each individual carries, as carrying more copies of such an allele is associated with an increased risk of developing Alzheimer's disease. The dataset also contains the p-values for the 30 SNPs with genome-wide significance in predicting Alzheimer's disease. These p-values are less than a predetermined threshold of 5E-8, and their corresponding SNPs can be determined using a Manhattan plot, which has gene loci on the horizontal axis and the negative logarithm with a base of 10 of the p-value on the vertical axis.

We will examine and use two methodologies applied to generate integrated risk models: LASSO and Ridge regression. LASSO and Ridge regression are two regularized generalized linear model approaches that address the bias-variance tradeoff often associated with fitting regression models. When including more predictors in a regression model, the bias of the model decreases while its variance increases. The inverse occurs when you reduce the number of the predictors, and this can result in a model that does not sufficiently or effectively explain the observed data. Regression techniques such as Ridge regression and LASSO perform regularization, which reduces the variance and shrinks values towards zero that are centered around zero in order to prevent the model from over-fitting the data.

LASSO, which stands for least absolute shrinkage and selection operator, is given by the following objective function:

$$Minimize: \sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

It is based on the following linear regression model:

$$Y_i = X_{ij}\beta_j + \epsilon_i$$

In this model, each Y represents an outcome, each beta value represents a coefficient, and each X represents the observed value of a predictor. The error term is normally distributed with mean zero and some variance greater than zero. The main focus for LASSO concerns the value of lambda, which is the regularization or tuning parameter. A larger tuning parameter increases the amount of regularization. Based on previous literature on LASSO regarding polygenic risk scores, we will set the tuning parameter equal to 0.001 in our model.

The second approach being used is Ridge regression, which is similar to LASSO, but it does not set any of the coefficients equal to zero, rather only shrinking them towards zero. Ridge regression minimizes the following objective function:

$$Minimize: \sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p}(\beta_j)^2$$

There is another regularization approach that combines elements of LASSO and Ridge regression known as Elastic Net, but this project focuses solely on comparing LASSO and Ridge regression. Elastic Net regression includes a parameter known as alpha, which equals one for LASSO and zero for Ridge regression, combining elements of the two approaches.

Using machine learning techniques, I divided the data into a testing set and a training set. The training set will be used to generate predictions for the affection status of the individuals, then the results will be compared to that of the testing set to determine the goodness of fit and accuracy of the models in terms of what was observed and what we expected. I created multiple training and testing sets of different sizes, with the ten principal component values, age, sex, and APOE status as predictors and affection status as the outcome. These specific predictors were chosen because the first ten principal components explain a large portion of the variability in the data, and age, sex, and APOE status are associated with the outcome of interest, which in this case is affection status (specifically developing Alzheimer's disease). With this selection of covariates, we will therefore be examining each method's ability to accurately generate a polygenic risk score with added covariates of age and sex (an integrated risk model).

Using the values of the above listed covariates, I generated a value for the coefficient for both the LASSO and Ridge regression models. Each coefficient is generated using the solveGlmnet function, which takes as inputs the predictor and outcome values, the value of the tuning parameter lambda, and the value of alpha, the other tuning parameter, which is different between LASSO and Ridge regression. It equals 1 for the LASSO model and 0 for the Ridge regression model. This coefficient is then multiplied by the values of the covariates from the testing test to obtain the expected value of our outcome of affection status. Using the norm function in R, I then obtain a norm of the difference between two matricies, one of which contains the expected values of our outcome based on the training set data and the other of which contains the values of our outcome based on the testing set data. This norm value provides a numerical value for how large the elements in the matrix are, though this is not dependent on the number of rows and columns of the matrix. This process is repeated for sample sizes 1000 through 2000.

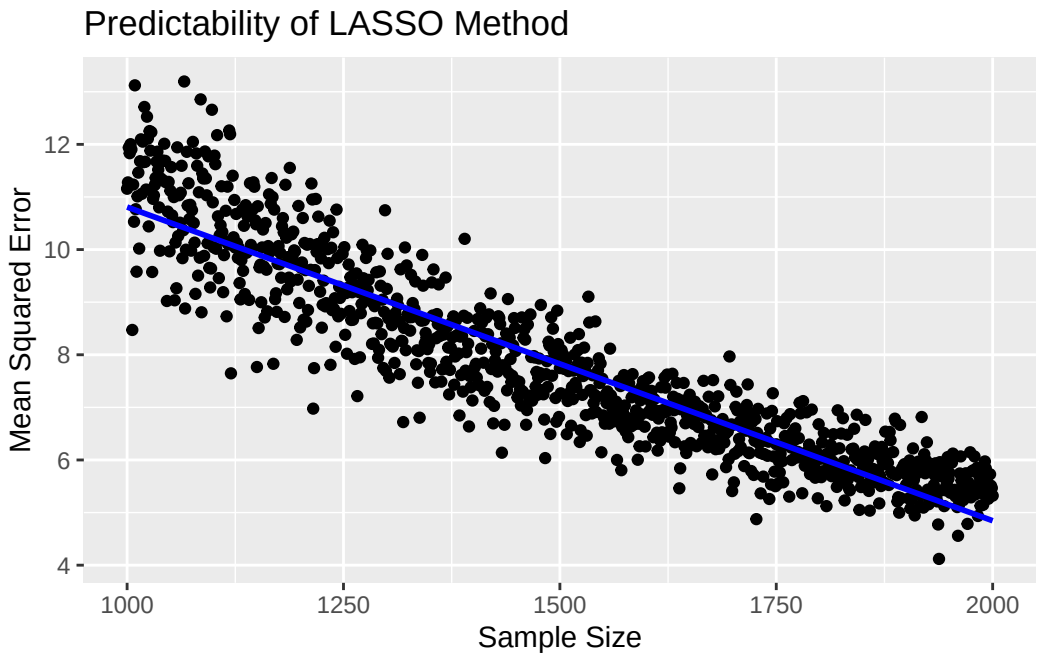
The norm values generated using the different methods will be evaluated in the following section to address the question of interest. We will also examine the mean squared error (MSE), which quantifies the amount of error in each model. The MSE will be calculated by dividing the obtained norm value by the sample size, and a lower MSE is desirable. It is also important to note that for this project, I used the following packages in R: Matrix, data.table,

dplyr, tidyr, qqman, glment, and ggplot2. A seed value of 100 was also set in order to ensure reproducibility of the results obtained. The dataset is not included in the submission of this project due to privacy restrictions.

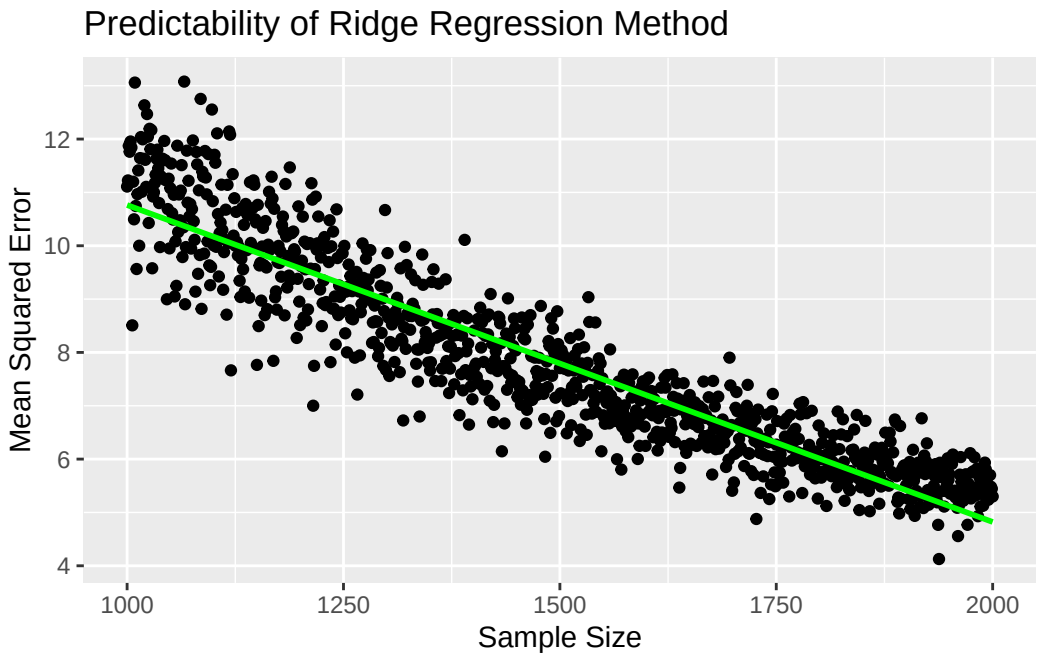
Results

In order to determine which methodology should be used to generate an intergrated risk model, I created two plots: one for LASSO and one for Ridge regression, shown in that order below. The plots have the sample size, corresponding to the size of the training set, on the horizontal axis and the value of the mean squared error for the respective sample size, calculate by dividing the norm value of the difference of the outcome matricies generated using the testing and training data by the sample size, on the vertical axis. As the size of the training set increases, we want to see the mean squared error decrease, as we are using more observed data to generate our integrated risk model. A linear regression line was added to each plot to show the extent of the linear relationship between the sample size and mean squared error.

```
`geom_smooth()` using formula = 'y ~ x'
```




```
`geom_smooth()` using formula = 'y ~ x'
```



Upon visual inspection, the two methodologies show a similar spread in terms of the data points seen in each plot. The linear regression lines also appear very similar in their slope values, which could suggest that the LASSO and Ridge regression models are quite similar in their production of integrated risk models and therefore polygenic risk scores (given the high amount of significance of the principal components and APOE information, all of which are determined via an individual's genetics).

To further examine the accuracy of the predictions that would be generated using each method, we examine the mean squared error for each method at the same sample size. A table of the ten randomly selected mean squared error values for each method is shown below. Each row corresponds to a different sample size.

	LASSO	Ridge
1	11.672109	11.631474
2	6.214975	6.196833
3	8.595886	8.575896
4	5.157362	5.143873
5	6.817581	6.805719
6	7.631836	7.588616
7	6.695041	6.693428
8	7.828342	7.842488
9	8.508009	8.493289
10	10.018979	9.999949

Although this is only a small sample of the 1001 mean squared error values generated, the Ridge regression method generates smaller mean squared error values at the respective sample size in nine of the ten instances. This indicates that Ridge regression may generate more accurate integrated risk models, as there is a smaller amount of error between the actual and predicted values for the outcome of affection status.

We can also examine the extent of the difference between the mean squared errors generated by each method for each sample size. The difference between the mean squared errors from each method for a respective sample size is shown in the table below, with the mean squared error from the Ridge regression method being subtracted from that of the LASSO method.

```
   Difference
1     0.04064
2     0.01814
3     0.01999
4     0.01349
5     0.01186
6     0.04322
7     0.00161
8    -0.01415
9     0.01472
10    0.01903
```

Because nine out of the ten values are positive, this shows that on average, the Ridge regression method will generate integrated risk models with outcomes that deviate less from the actual outcome than when the LASSO method is used. When examining all of the 1001 differences in mean squared error values between the two methods, 982 out of the 1001 were positive, suggesting that the Ridge regression method provides a more accurate integrated risk model than the LASSO method 98.1% of the time.

To test whether or not this difference is statistically significant, we can perform a t-test, assuming that the data satisfy the assumptions of normality and homoscedasticity. A high p-value would suggest that the difference is not statistically significant.

```
    Welch Two Sample t-test

data:  df_Lasso_MSE and df_Ridge_MSE
t = 0.40906, df = 1999.9, p-value = 0.6825
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1293735  0.1975668
sample estimates:
```

```
mean of x mean of y
 7.826425  7.792328
```

As seen above, with a p-value of 0.6825, we have statistically significant evidence that the difference in the mean squared error generated at each sample size is not different from zero. This suggests that although the Ridge regression method may perform slightly better on average, it does not perform significantly better than the LASSO regression method.

The receiver operator characteristic, or ROC, curve can also be used to quantify the performance of certain models, plotting specificity and sensitivity. The closer to 1 the area under the curve is, the fewer the number of false positives generated by the model used to generate the ROC curve. For the sake of brevity, these ROC curves will not be included, but the generation of such curves will build upon the results obtained through this project.

Based on these results, we can conclude that the Ridge regression and LASSO methods are the extremely similar in terms of their ability to generate an integrated risk model for Alzheimer's disease, but the Ridge regression method performs slightly better in terms of the accuracy of the integrated risk model generated.

Conclusion

My project aimed to determine which methodology would generate the best integrated risk model for Alzheimer's disease, using machine learning techniques to create a training set and testing set to evaluate a LASSO model and a Ridge regression model. Based on the results above, we conclude that the Ridge regression method is slightly better in terms of generating an integrated risk model that is most predictive of an individual's likelihood of developing Alzheimer's disease. It is important to note that the extent to which Ridge performs better than LASSO is not statistically significant.

In terms of pitfalls of my project, it would be ideal to test a larger number of methodologies and more complex methodologies that are currently being worked on by researchers in the field, though I am limited in multiple ways. I am computationally limited by the device I use to conduct my research, though these issues arise more so when the methodology becomes more complex and when the size of the dataset increases significantly, the latter of which I have yet to encounter. It can also be quite difficult for all software to run properly on my device, which limits being able to use of various methodologies that require such software to generate polygenic risk scores. Another limitation concerns the data used for our testing and training sets. All of the individuals were non-Hispanic White individuals, which may mean that the results we obtain are not applicable to individuals who do not have this ethnicity. To build upon my project, future research could be conducted using data from individuals across many ethnicities to evaluate whether or not the results from above still hold. Other projects could determine how to choose a threshold value for calculating the area under the ROC curve that minimizes error, and whether or not doing so provides the same or different results

from the results obtained without using such information. Future projects could also compare the methodologies examined in this paper to other methodologies used to generate polygenic risk scores, such as MegaPRS and LDAK. The end goal is determining which methodology generates a polygenic risk score and integrated risk model with the most predictive power, a field with increasing attention and new developments constantly arising.

References

Data Science Labs. (2023). Regularization. https://datasciencelabs.github.io/2023/29-regularization.html

Dalmasso, M. C., Brusco, L. I., Olivar, N., Muchnik, C., Hanses, C., Milz, E., … & Bartesaghi, L. (2020). An integrative multi-omics analysis identifies epigenetic alterations associated with Alzheimer's disease. Alzheimer's Research & Therapy, 12(1), 1-17. https://alzres.biomedcentral.com/articles/10.1186/s13195-020-00740-0

Hahn, G., Prokopenko, D., Lutz, S. M., Mullin, K., Tanzi, R. E., Cho, M. H., Silverman, E. K., Lange, C., & on behalf of the NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium. (2022). A Smoothed Version of the Lassosum Penalty for Fitting Integrated Risk Models Using Summary Statistics or Individual-Level Data. Genes, 13(1), 112. https://doi.org/10.3390/genes13010112

Jonas, S., Izarzugaza, J. M. G., Ried, T., & Kaderali, L. (2020). Detecting signatures of mutational processes operating in human cancer. Genome Medicine, 12(1), 1-11. https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-020-00742-5

Knowles, J. W., & Ashley, E. A. (2021). Cardiovascular disease: The rise of the genetic risk score. Nature Communications, 12(1), 1-3. https://www.nature.com/articles/s41467-021-24082-z

Lake, E. (2022). Topic 5: The Linear Model - Model Selection. BST 210 Applied Regression Analysis. (n.d.).

National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS). (n.d.). https://www.niagads.org/home

Torres, L., & Vigouroux, C. (2022). Challenges in identifying genetic variants associated with pharmacogenetic traits in African populations. Frontiers in Genetics, 13, 1-10. https://www.frontiersin.org/articles/10.3389/fgene.2022.818574/full