HACK-A-STAT 2025

Group Name :- StatLab Trio

Robin Pereira : A056

Rishabh Pandey : A041

Jainam Gada : A019

Problem Statement

A genetic study was conducted using 50 mice. During the study, for each mouse, gene expression data corresponding to 2000 genes was also collected. Along with gene expression data, data corresponding to a phenotype was also collected. The objective is to identify which genes have a significant impact on the phenotype. Create a presentation to answer this question of interest.

Objective

The objective is to identify which genes have a significant impact on the phenotype.

Introduction

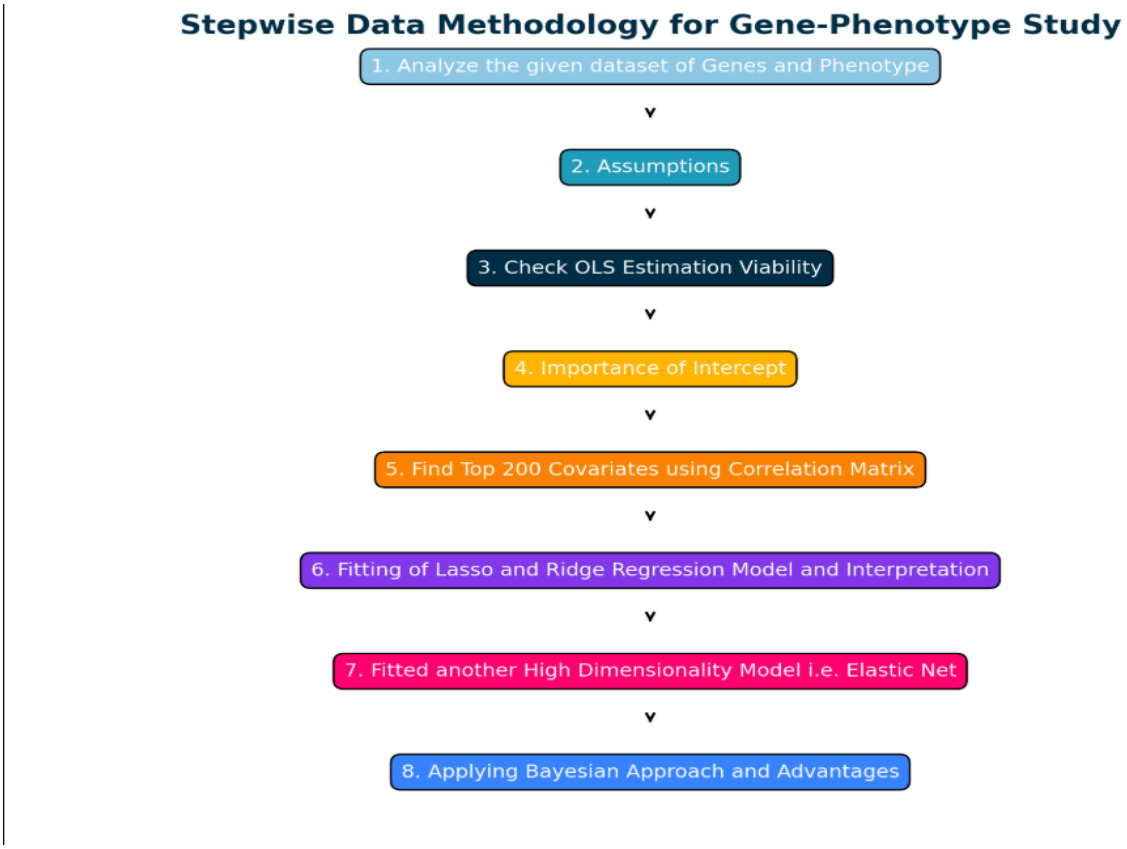
In recent years, progress in genetic research has helped scientists better understand how genes affect visible traits. A major task in this field is to find out which specific genes play an important role in these traits. This report describes a study done on 50 mice, aimed at finding out which genes might influence a certain trait by looking at gene activity.

In this study, data on the activity of 2000 different genes to each of the 50 mice was provided . Information about a specific visible trait, called a phenotype (target variable) had also been provided . By combining these two sets of data, our task was to find out which genes are closely linked to changes in the trait.

Finding these significant genes helps to identify how traits develop: It helps us learn about the biological processes behind the trait. It allows us to create models that predict traits based on genetic data and further provides clues for finding new treatment options or areas for more research.

This report will explain how we analyzed the data, describe the methods used to find significant genes, and show the results. We will also discuss the findings and suggest ideas for future research.

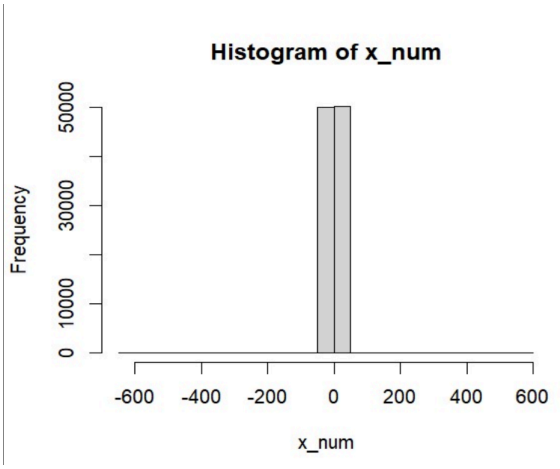
Methodology



We are given the dataset of 2000 genes and Phenotypes of 50 mice.

For Gene Data :

- Mean: The average value across the gene dataset columns is approximately close to 0, indicating a balanced distribution around zero.



- Standard Deviation: The average standard deviation is approximately close to 1.00, suggesting moderate variability in the gene data.

This shows that gene data shows a balanced distribution with moderate variability, which is typical for datasets of this nature. The symmetric range of minimum and maximum values suggests no extreme outliers.

Assumptions for the Model

1. **Multicollinearity**

   Correlation Matrix: The correlation matrix of the features in genes indicates that there is no significant multicollinearity. The correlation coefficients are generally low, with means around zero and standard deviations of 0.14 to 0.15, suggesting that the features are not highly correlated with each other.

2. **Normality of Target Variable**

   Shapiro-Wilk Test: The Shapiro-Wilk test for normality on the target variable resulted in a test statistic of 0.99 and a p-value of 0.89. This

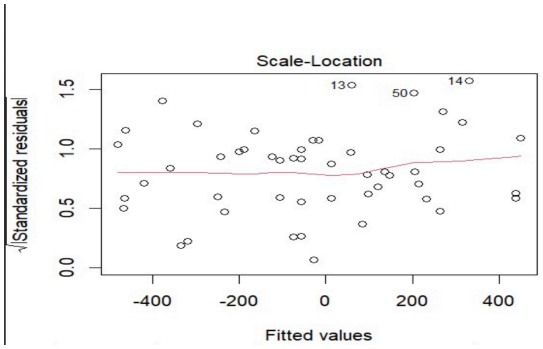
suggests that the target variable is normally distributed, as the p-value is greater than the common alpha level of 0.05.

3. Linearity

Correlation with Target: The correlation coefficients between the features and the target variable are generally low, with a mean of 0 and a standard deviation of 0.21.

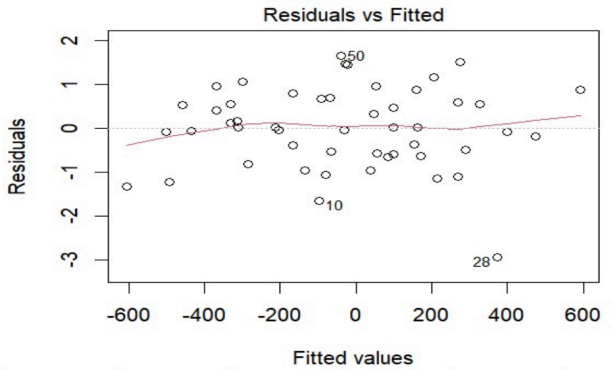
4. Homoscedasticity

Residuals vs Fitted Values: The residuals from the simple linear regression model show no apparent pattern when plotted against the fitted values, suggesting homoscedasticity.



5. Independence of Errors

Residual Analysis : The residuals from the regression model show no discernible pattern, with a mean of 0 and a standard deviation of 1. This suggests that the errors are independent.



OLS Estimation Viability

Our data is a high dimensional data. In this dataset we have our genes (predictors) more than our number of observations.

This would mean that our model is not a full rank model. As the model is not full rank we will not be able to get unique coefficient estimates since X'X is not invertible. This would mean that to find our values of the coefficients we will have to use the g-inverse of X'X which is not unique.

Thus this would mean that a crucial assumption of OLS is **not** satisfied and hence OLS estimation would not be feasible in our case.

Having 2000 gene(predictor) data would mean that there would be too many coefficients to interpret and complex interactions would not be captured due to the presence of correlation between the various genes and that would undermine our objective of finding specific genes that significantly impact phenotype.

Importance of Intercept

Yes the intercept is required in the model because In this particular dataset, intercept means the value of the phenotype when the values of the genes are zero. If we don't include intercept in the model that would mean that the value of the phenotype(Target variable) is also zero when the value of genes is zero.

This does not show the actual relationship of phenotypes and the genes as this forces the regression line to pass through the origin.

Use of Univariate filtering

- Using the Correlation based Feature selection , we computed the correlation coefficient between each predictor and the target variable .
- Then Ranked the predictors based on the absolute value of their correlation coefficients.
- Then the top 200 predictors with the highest absolute correlation values were selected .
- Thus Correlation-based feature selection provides a simple yet effective way to reduce the number of predictors.
- Below are the 20 covariates with the highest correlation :-

Fitting Ridge and Lasso Regression Models

Ridge Regression :-

- Ridge Regression is a L1 regularization technique that works by helping reduce the potential for overfitting to the training data. Ridge Regression performs the best when a lot of variables contribute equally though quite less to the model.
- The goal of Ridge Regression is to help prevent overfitting by adding an additional penalty term .

$$RSS = \sum_{i=1}^{n} (yi - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

- Ridge Regression is used to minimize this entire error term RSS + Penalty.
- Where λ is a hyperparameter which is used to tune the model
- λ determines how serve the penalty is it can be between 0 to ∞
- In our analysis we have used Ridge Regression with the value of λ as 0.01 using

The test train method where we have divided the data into two parts one i.e. 0.2 for testing and 0.8 for training the data

The Findings that we got for ridge regression is

Model	R^2	MSE	Standard deviation of residuals.	Expected variance score
Ridge Regression	0.9302	8646.7006	90.2987	0.934179

Lasso Regression :-

Lasso regression is a linear regression technique that incorporates L-2 regularization and feature selection .

Lasso can force some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large.

The Lasso performs variable selection similar to subset selection. Such models are easier to interpret. Lasso models performs the best when few variables in the dataset contribute significantly to the model

Model	R^2	MSE	Standard deviation of residuals.	Expected variance score
Lasso Regression	0.957	5250.175382	70.8466	0.959483

Comparison of Ridge and Lasso Regression.

Aspect	**Lasso Regression**	**Ridge Regression**
Effect on Coefficients	Shrinks some coefficients to exactly zero, effectively performing feature selection.	Shrinks coefficients towards zero but does not make them exactly zero.
Feature Selection	Yes, Lasso can eliminate irrelevant features by setting their coefficients to zero.	No, Ridge includes all features, but shrinks their influence.
R^2	0.957	0.9302
MSE	5250.175382	8646.7006
Standard deviation of residuals	90.2987	70.8466
Expected Variance S	0.959483	0.934179

As we can see that the R^2 and MSE of lasso is better than Ridge Regression so we can conclude that the L 2 regularization i.e. lasso is better in capturing the information provided by the independent variable than Ridge Regression.

Elastic Net Regression:

Elastic net regression is a regularized regression method that combines L1 (Lasso regression) and L2 (Ridge regression) penalties in its objective function.

For a linear regression model with response variable Y and predictors X, the elastic net estimator β minimizes:

$$||Y - X\beta||^2 + \lambda[(1-\alpha)||\beta||^2 + \alpha||\beta||_1]$$

Where:

- $||Y - X\beta||^2$ is the regular sum of squared residuals
- $||\beta||^2$ is the L2 (ridge) penalty = $\Sigma\beta j^2$
- $||\beta||_1$ is the L1 (lasso) penalty = $\Sigma|\beta j|$

- $\lambda \geq 0$ is the overall regularization strength
- $\alpha \in [0,1]$ is the mixing parameter between L1 and L2 penalties

Variable Selection: Automatically selects important features and eliminates irrelevant ones by shrinking coefficients to zero.

Oracle Properties: Gets better at identifying truly important variables as sample size increases.

Bias-Variance Trade-off: Introduces a small bias to achieve more stable predictions across different samples.

Special Cases: Works as ridge regression ($\alpha=0$), lasso regression ($\alpha=1$), or a customized blend of both ($0<\alpha<1$).

Cross-Validation: Uses grid search to find optimal values of λ (penalty strength) and α (mixing parameter).

Model	R^2	MSE	Standard deviation of residuals.	Expected variance score
Elastic Net Regression	0. 9576	5250.175382	70.8466	0.959483

As we can see that the MSE and R^2 of elastic net and lasso are very close.From the findings we can say that the regularization and feature selection properties of Lasso are effectively capturing the relevant predictors.

Bayesian Approach for Ridge Regression and the Lasso Regression

Let's view ridge regression and the lasso regression through Bayesian lens. In high-dimensional data , Bayesian approaches help in regularization and variable selection.A Bayesian viewpoint for regression assumes that the coefficient vector β has some prior distribution.

Therefore, from a Bayesian viewpoint, ridge regression and the lasso follow directly from assuming the usual linear model with normal errors, together with a simple prior distribution for β. Notice that the lasso prior is steeply peaked at zero, while the Gaussian is flatter and fatter at zero. Hence, the lasso expects a priori that many of the coefficients are (exactly) zero, while ridge assumes the coefficients are randomly distributed about zero.

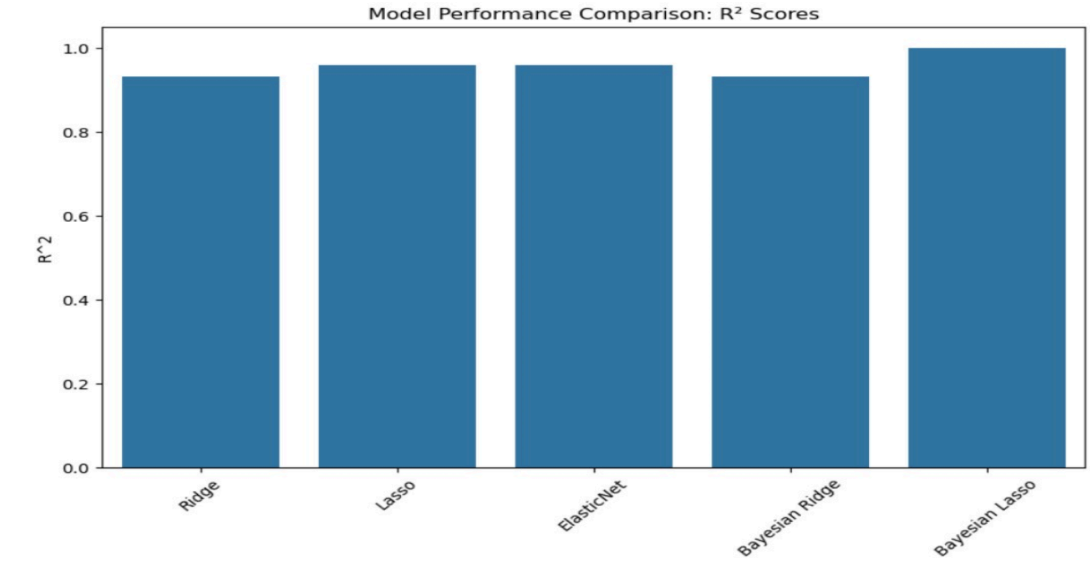
Surely there is a benefit in taking a Bayesian approach to high dimensional regression methods.

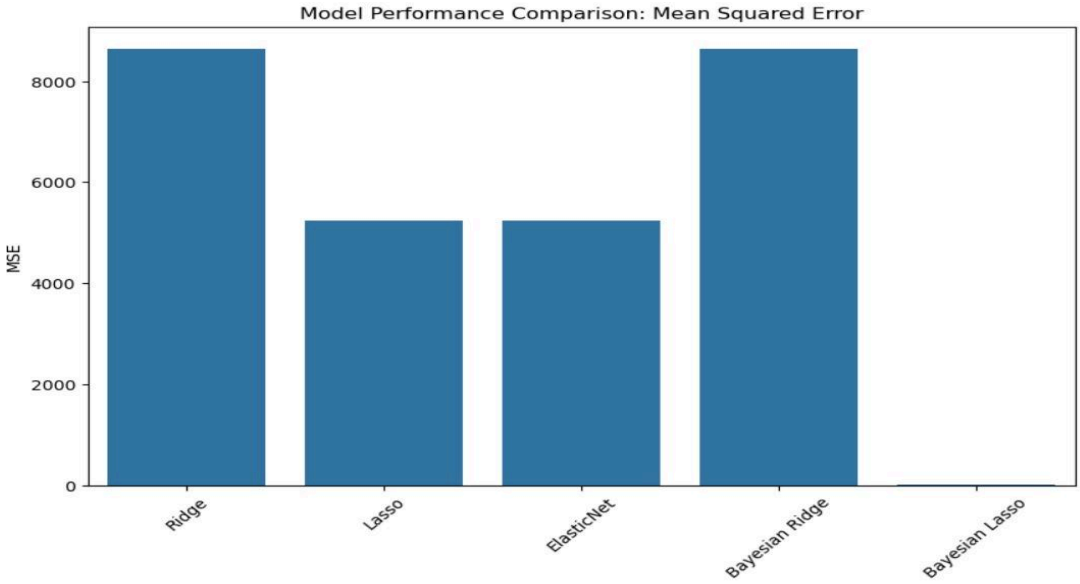
Model Evaluation

Model	R^2	MSE	Standard deviation of residuals.	Expected variance score
Ridge	0.930206	8646.033091	70.8466	0.934179
Lasso	0.957619	5250.175382	90.2987	0.959483
Elastic Net	0.957619	5250.175382	70.8466	0.959483
Bayesian Ridge	0.930207	8645.958930	90.2983	0.934180
Bayesian Lasso	0.999926	9.218182	3.0245	0.959483

We have compared the models with respect to different performance evaluation metrics and our findings suggest that the ***Bayesian Lasso model*** out performs every other model in terms of R^2 , MSE ,Standard deviation of residuals and Expected variance score.

From the above finding we can say that the bayesian lasso is the best model for dealing with high dimensionality data which has two many independent variables .





From the above graph we can see that Bayesian Lasso has a minimum MSE.

Conclusion

We have seen from the above analysis that Bayesian Lasso Regression is the best fit for the model. Since Bayesian Lasso seems to be a good fit below we have taken the absolute coefficients from the Bayesian Lasso model and based on these

coefficients we will find out the Genes having the most significant impact on Phenotype.

Genes	Absolute Coefficients
X2	**177.894748**
X1	**131.015314**
X5	**88.704053**
X4	**76.764223**
X3	**47.186593**
X273	**0.572452**
X713	**0.489550**
X142	**0.429967**
X1628	**0.346750**
X472	**0.339163**

The main objective of our project was to identify the independent variables (genes) that significantly impact our dependent variable (phenotype) from a set of 2000 variables.

Using the model (Bayesian Lasso) and their coefficients, we identified the top 10 Genes that have the most substantial influence on the Phenotype.