HACK-A-STAT

Group Name: Analytica

Group Members

Name	Roll No	SAP ID
Shruti Khapekar	A028	86062400066
Sourabh Lad	A030	86062400017
Sandesh Tayde	A070	86062400020

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. You may use the following hints to create your presentation:

1. For this dataset is it feasible to use Ordinary Least Square estimation?
2. Is an intercept required in the model? Justify your answer and proceed accordingly.
3. Find a simple screening method to reduce the number of covariates to 200.
4. Can a high dimensional regression approach be used here to analyze this dataset? What are the assumptions under which these methods can be utilized?
5. Implement popular high dimensional regression approaches like Ridge and Lasso. Compare both methods and comment on which works better and why.
6. Is there a benefit in taking a Bayesian approach to high dimensional regression methods mentioned above? Try implementing the Bayesian approach and comment if you observe any advantages.

Dataset Summary:

Genotype Data

Dimensions: 50 rows (mice) × 2000 columns (genes)

Contains gene expression levels for each mouse.

Phenotype Data

Dimensions: 50 rows (mice) × 1 column

Contains the phenotype value for each mouse.

Objective:

1. Determining the gene that significantly impact the phenotype of the mice.
2. Implementing screening method to reduce number of covariates (genes).
3. Comparing the performance of Ridge and Lasso regression and choosing the best method to identify important genes.
4. Use Bayesian regression method for identifying significant genes and compared it with ridge and lasso regression method.

Ordinary Least Squares (OLS) Method

Ordinary Least Squares is a statistical method used to estimate the relationship between a dependent variable and one or more independent variables. The goal is to identify the best-fitting line, represented by coefficients, that explains the relationship between the variables.
OLS assumes linearity, independence, homoscedasticity, and no multicollinearity among the predictors.

Mainly focusing on two aspects:

1. Mean Squared Error (MSE)
2. Multicollinearity

1. Mean Squared Error (MSE)

The MSE measures the average of the squared differences between predicted values and actual target values. By squaring the differences, the MSE places a higher weight on larger errors, making it sensitive to outliers. A lower MSE indicates that the model's predictions are closer to the true values, reflecting better overall performance.

2. Multicollinearity

For multicollinearity we will be using condition index. The Condition Index relies on the eigenvalue decomposition of the correlation matrix of predictor variables.

Eigenvalues: Represent the variances of different linear combinations of the predictors. **Eigenvectors:** Represent the directions of these linear combinations.

Condition Number: The largest Condition Index is the "condition number," calculated as:

$$\text{Condition Number} = \sqrt{(\text{Maximum Eigenvalue} / \text{Minimum Eigenvalue})}$$

This ratio reflects how "spread out" the eigenvalues are. If the ratio is large, it means some linear combinations of predictors have much larger variances than others.

Interpretation:

- **Low Condition Index (close to 1):** Minimal multicollinearity. Regression coefficients are stable and reliable.
- **Moderate Condition Index (between 10 and 30):** Moderate multicollinearity. Coefficients might be unstable and require careful interpretation.
- **High Condition Index (above 30):** Severe multicollinearity. Coefficients are highly unstable and unreliable. The model's predictions may be inaccurate and misleading.

Analysis:

1. Mean Squared Error (MSE)

Training MSE	3.443919393640868e-25
Testing MSE	101898.66106339319

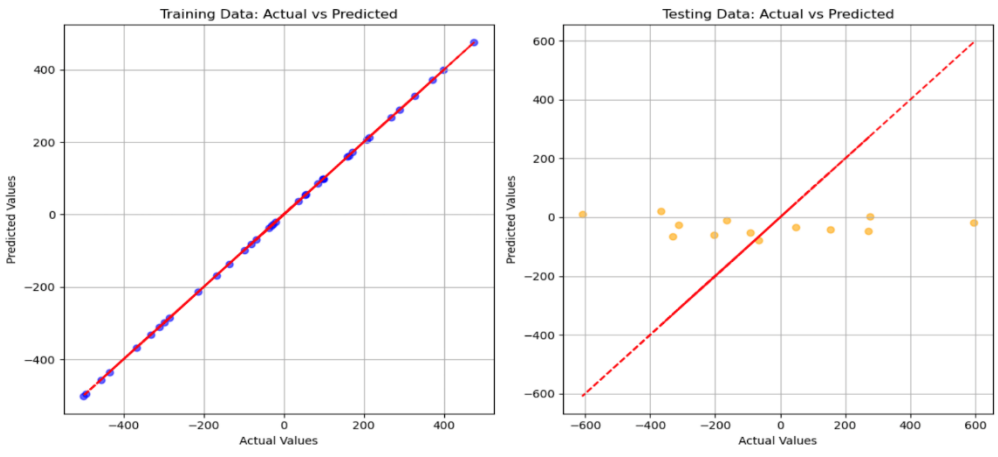
The OLS model achieved an extremely low Training MSE, which suggests that the model fits the training data very closely. However, the much higher Testing MSE of 101,898.66 indicates that the model does not generalize well to unseen data. The significant gap between the Training and Test MSE suggests overfitting. Given this, OLS is not suitable for this dataset.

2. Condition Number

Condition Number (Index)	812676664997782.2

The condition number of 812676664997782.2 suggests severe multicollinearity in the dataset, as values above 30 indicate instability in regression coefficients. This makes OLS regression unreliable for estimating relationships between genes and phenotype.

When we try to fit the model the following result, we get i.e. Overfitted model



Intercept Study

Using Ridge Regression

Root Mean Squared Error (Full Model): 0.9300365655664293

Root Mean Squared Error (Restricted Model): 0.9300365655664291

The restricted model (without intercept) performs similarly. Intercept may not be necessary.

Using Maximum likelihood estimator

Root Mean Squared Error (Full Model): 0.001444

Root Mean Squared Error (Restricted Model): 0.001421

The ability of the model to perform on unseen data is not affected by the value of the intercept, i.e. the intercept does not make a significant impact to the model.

Conclusion:

Intercept is nothing but the value of the response variable when all the regressor are zero. In this case the value of all the genotype at the same time point cannot be equal to zero, hence intercepting the value of phenotype without having knowledge of at least one genotype is not possible. Therefore, intercept does not contribute to the model significantly.

High Dimensional Regression Approach

The high-dimensional regression methods can be applied to the dataset with 50 observations (mice) and 2000 predictors (genes) which is latter converted into 200 features using simple feature screening method to predict the phenotype trait.

The following methods are designed to handle situations where the number of features-Mice greatly exceeds the number of samples (p>n) i.e. the number of features is greater than the cases.

Ridge Regression:

- Ridge regression, also known as L2 regularization, utilizes a penalty comprises the tuning parameter multiplied by the squared sum of the coefficient values.
- Handles multicollinearity effectively by shrinking coefficients, but does not perform feature selection.

Lasso Regression:

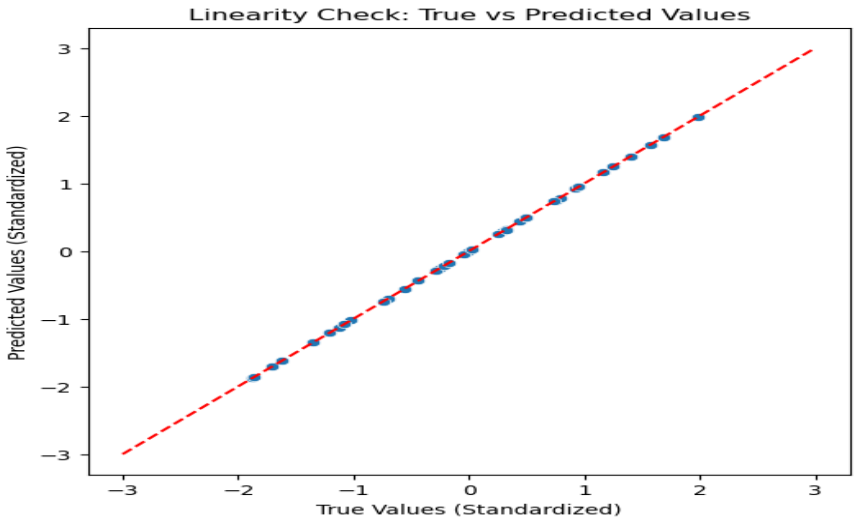
- In Lasso regression, also known as L1 regularization, a penalty term is added to the coefficients that is proportional to their absolute values.
- Performs feature selection by removing irrelevant features.

Elastic Net:

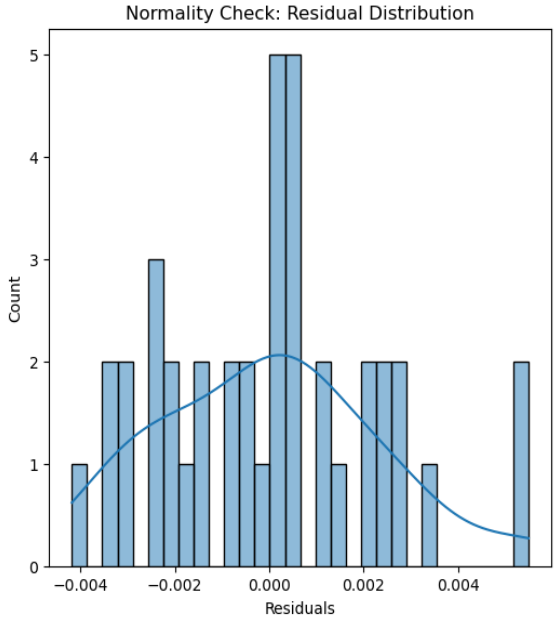
- Elastic Net regression combines the strengths of Ridge and Lasso regression, which involves both penalties $L1$ and $L2$.
- Provide a flexible approach to feature selection and regularization in datasets with multicollinearity or when the number of predictors exceeds the number of observations.

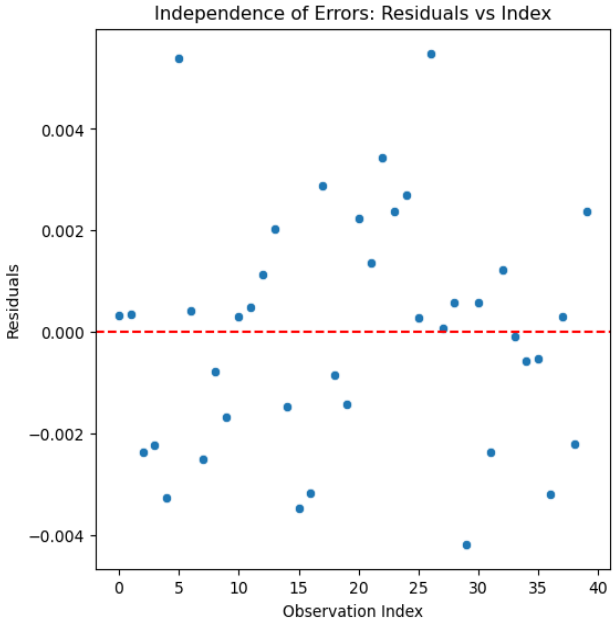
Assumptions:

- **Linearity**: The points lie very close to the red dashed diagonal line; The assumption of linearity is likely satisfied.

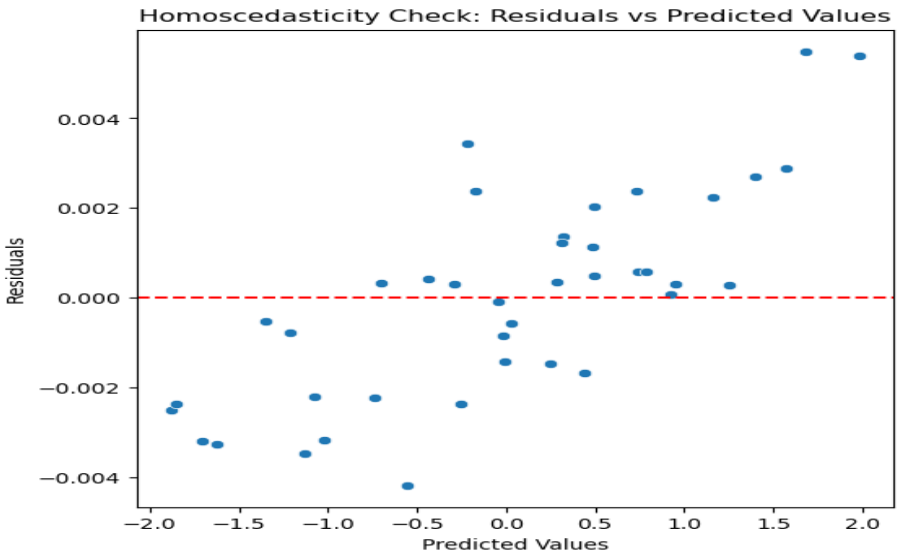


- **Independence of Errors**: Errors should be uncorrelated and normally distributed.





- **Noise Free Data:** The data should be clean and noise-free.
- **Homoscedasticity:** Variance of the residuals is constant across all levels of the independent variables.



Interpretation:

All necessary assumptions about the high-dimensional regression methods are satisfied.

Covariate Reduction & Model fitting

There are various methods to reduce the data but due to very dimension they cannot be applied. One of the most commonly used method is PCA (Principal Component Analysis), but in this case the number of variables is very large than size of the dataset. Range of features reduction is from 0 to 50 (in this case), we want reduction up to 200 covariates; so, the methods fail. Other methods are discussed below.

1. Correlation Analysis for Feature Selection

Firstly, we applied correlation to get the 200 covariates which were more important to the model. The variates with high correlation were preferred.

Now, considering the new data set which will have the 200 genes is considered for further testing and model building.

a) Ridge Regression

For Testing data set:

Mean squared error	Root Mean squared error	R2_score
0.1144	0.33827	0.9113

For Training data set:

Mean squared error	Root Mean squared error	R2_score
4.6684 e-06	0.00216	0.9999

Ridge Regression (With Hyperparameter Tuning)

For Testing data set:

Mean squared error	Root Mean squared error	R2_score
0.1137	0.3372	0.9118

For Training data set:

Mean squared error	Root Mean squared error	R2_score
4.7159 e-08	0.000217	0.9999

Interpretation:

- RMSE of 0.33 is the square root of the MSE indicates good model accuracy.

- A very low RMSE on the training set suggests minimal error in predictions, reinforcing the possibility of overfitting unless the data is exceptionally clean and noise-free.
- An R2 score of 0.911 means that your model explains approximately 91.1% of the variance in the test data. This is quite high, indicating that your model fits the test data well.
- An R2 score of 0.99 on the training set indicates a almost perfect fit, where your model perfectly predicts the target variable based on the training data. However, this could also indicate overfitting, especially if the test R2 score is significantly lower.

b) Lasso Regression

For Testing data set:

Mean squared error	Root Mean squared error	R2_score
1.3116	1.1452	-0.01654

For Training data set:

Mean squared error	Root Mean squared error	R2_score
0.89244	0.9446	0.0

R2_score is 0 indicating that no variation is explained by the model.

c) Elastic Net Regression

For Testing data set:

Mean squared error	Root Mean squared error	R2_score
1.1419	1.0686	0.1149

For Training data set:

Mean squared error	Root Mean squared error	R2_score
0.74104	0.8608	0.1696

Interpretation:

- RMSE of close to 1, indicates not a good model.
- An R2 score of 0.1 means that your model explains approximately 10% of the variance in the data so overall it is not a good model.

Comparison between Ridge and Lasso Regression:

Ridge model gives more accuracy in results as well as it explains model well as compared to others.

d) Bayesian Ridge Regression:

For Testing data set:

Mean squared error	Root Mean squared error	R2_score
0.1136	0.3371	0.9112

For Training data set:

Mean squared error	Root Mean squared error	R2_score
7.11306 e-12	2.6703e-06	0.9999

Interpretation:

- RMSE of model is low i.e. the square root of the MSE indicates good model accuracy.
- A very low RMSE on the training set suggests minimal error in predictions, reinforcing the possibility of overfitting unless the data is exceptionally clean and noise-free.
- An R2 score of 0.91 means that your model explains approximately 91% of the variance in the test data. This is quite high, indicating that your model fits the test data well.
- An R2 score of 0.99 on the training set indicates a almost perfect fit, where your model perfectly predicts the target variable based on the training data. However, this could also indicate overfitting, especially if the test R2 score is significantly lower.

Comment:

The Ridge Regression is performing better than the Lasso Regression Because, there is multicollinearity in the data and Ridge Regression handles multicollinearity effectively by shrinking the coefficients.

2. K Feature Selection Method

K-Feature Selection selects the K most important features from a dataset to improve model performance and reduce complexity. Features are evaluated, ranked, and the top K are chosen based on their relevance to the target variable.

a) Ridge Regression

For Testing data set:

Mean squared error	Root Mean squared error	R2_score
0.1144	0.3382	0.9113

For Training data set:

Mean squared error	Root Mean squared error	R2_score
4.6684 e-06	0.0021	0.9999

Interpretation:

- RMSE of model is low i.e. the square root of the MSE indicates good model accuracy.
- A very low RMSE on the training set suggests minimal error in predictions, reinforcing the possibility of overfitting unless the data is exceptionally clean and noise-free.
- An R2 score of 0.91 means that your model explains approximately 91% of the variance in the test data. This is quite high, indicating that your model fits the test data well.
- An R2 score of 0.99 on the training set indicates a almost perfect fit, where your model perfectly predicts the target variable based on the training data. However, this could also indicate overfitting, especially if the test R2 score is significantly lower.

b) Bayesian Ridge Regression

For Testing data set:

Mean squared error	Root Mean squared error	R2_score
0.1136	0.3371	0.91192

For Training data set:

Mean squared error	Root Mean squared error	R2_score
7.1306 e-12	2.67033	0.9999

Interpretation:

- RMSE is high indicating overfitting.
- An R2 score of 0.91 means that your model explains approximately 91% of the variance in the test data. This is quite high, indicating that your model fits the test data well.
- An R2 score of 0.99 on the training set indicates a almost perfect fit, where your model perfectly predicts the target variable based on the training data. However, this could also indicate overfitting, especially if the test R2 score is significantly lower.

3. Recursive Feature Elimination

RFE is a feature selection technique that identifies the most important features for a model by recursively removing the least important ones.

a) Ridge Regression

For Testing data set:

Mean squared error	Root Mean squared error	R2_score
0.1225	0.35007	0.9050

For Training data set:

Mean squared error	Root Mean squared error	R2_score
2.7319 e-08	0.00016	0.9999

Interpretation:

- After tunning the model, RMSE for testing data is more than that for testing this is caused due to overfitting.
- The R2_score is 0.99 for testing data, indicating a good model.

b) Bayesian Ridge Regression

For Testing data set:

Mean squared error	Root Mean squared error	R2_score
0.1224	0.3499	0.9051

For Training data set:

Mean squared error	Root Mean squared error	R2_score
3.1124 e-24	1.7621e-12	1.0

Interpretation:

- After tunning the model, RMSE for testing data is more than that for testing this is caused due to overfitting.
- The R2_score is 0.99 for testing data, indicating a good model.

Conclusion:

We observed that even after applying Ridge regression there is overfitting in the model. To remove overfitting, we calculated the VIF (Variance inflation factor). We have set the threshold to be equal to 10 because we are using Ridge Regression.

Advantages of Bayesian Ridge Regression over Ridge Regression:

- The model automatically decides the amount of regularization needed.
- It works well when many genes are highly related to each other, helping the model stay stable.
- The model provides more detailed information about the importance of genes, making it easier to interpret.
- It works well when you have many genes (features) and fewer samples (data points), which can be tricky for other models

Variance Inflation Factor (VIF):

Variance Inflation Factor (VIF) tells us about the correlation between independent variables irrespective of the response variable.

VIF = 1: No multicollinearity.
$1 < VIF \leq 5$: Moderate multicollinearity.
VIF > 10: Severe multicollinearity, which can impact the model's reliability.

We observed that even after applying Ridge regression there is overfitting & multicollinearity in the model. To remove multicollinearity, we calculated the VIF (Variance inflation factor) and removed the covariates above a given threshold value. We have set the threshold to be equal to 10 because we are using Ridge Regression.

We obtained 44 covariates whose VIF was more than threshold value.

After performing the analysis again over all the three methods mentioned above there was overfitting in the model.

Stacked Regression:

To overcome the problem of overfitting we made use of Stack Regression with base learners as Regression and Decision Tree.

1. Correlation Analysis for Feature Selection

We took the same data as used above in case of Correlation Analysis for Feature Selection.

Hyperparameter (Best Ridge alpha): 0.001

Training data:

R^2	0.9538
RMSE	0.2030

Testing data:

R^2	0.9443
RMSE	0.0719

The RMSE value is low for both the training as well as testing data.

2. Recursive Feature Elimination

We took the same data as used above in case of Recursive Feature Elimination.

Hyperparameter (Best Ridge alpha): 0.001

Training data:

R^2	0.8593
RMSE	0.3543

Testing data:

R^2	0.9807
RMSE	0.0248

The RMSE value is low for both the training as well as testing data.

3. K Feature Selection Method

We took the same data as used above in case of K Feature Selection Method.

Hyperparameter (Best Ridge alpha): 0.001

Training data:

R^2	0.8593
RMSE	0.3543

Testing data:

R^2	0.9807
RMSE	0.0248

The RMSE value is low for both the training as well as testing data.

Results and Conclusion

Results:

1. Due to high multicollinearity and overfitting ordinary least square method cannot be used.
2. Intercept does show any significance in the ridge regression model.
3. K Feature Selection Method, Recursive Feature Elimination, Correlation Analysis are the methods to reduce the no. of covariates to 200. The best results were obtained from Correlation Analysis.
4. The Ridge regression model was taken and not Lasso regression model because its works with multicollinear data well.
5. Bayesian Ridge Regression model automatically decides the amount of regularization and work good on overfitted data.
6. Stacked Regression also tells that Correlation analysis gives the best model.

Conclusion:

1. All the genotype at the same time point cannot be equal to zero, hence intercepting the value of phenotype without having knowledge of at least one genotype is not possible.
2. Overfitting occurs because loss of information about even one gene will lead to change in the value of phenotype, which can mislead the result.
3. There is high multicollinearity i.e. the 2000 gene that are given depends on each others behaviour.
4. The feature selection of 200 covariates is nothing but the most important 200 genes which have correlation between them. They will affect the behaviour(value) of phenotype the most.
5. Bayesian Ridge Regression Ridge Regression model on an average 90% of times predicts the value of phenotype correctly when given the value of genotype.