Hack-a-Stat 2024
Group Name: Databloom

Payal Shah - A065
Drishti Shah - A064
Himani Grover - A020

Contents

1 Introduction

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 significantly impact the phenotype. Understanding the relationship between gene expression and phenotypes is critical in genetics and bioinformatics.

2 Objectives

- Identify significant genes impacting the phenotype using high-dimensional regression techniques.
- Determine the feasibility of using Ordinary Least Squares (OLS) regression.
- To determine the simple screening method to reduce the dimensionality of the data.
- Comparing high-dimensional regression approaches like Ridge, Lasso, and Bayesian regression.

3 Data Overview

- **Gene Expression Data:**
  Total samples: 50 (phenotype).
  Total features: 2000 (genes).
- **Phenotype Data:**
  Response variable: 50 (A measurable phenotype corresponding to each mouse).

4 Methodology

4.1 Ordinary Least Squares (OLS) Estimation

$$y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \epsilon \tag{1}$$

In order to fit a regression model to the data, we need to estimate the parameters. OLS is one of the best-known methods for parameter estimation. However, the data is high-dimensional. Hence, the model is not a full-rank model. Thus, $X'X$ inverse does not exist, and unique solutions cannot be found for the parameters. Hence, OLS is not feasible for estimating the parameters.

4.2 Need for Intercept

The requirement of an intercept in a regression model depends on the nature of the data and the assumptions that are made about the relationship between the independent variables and the dependent variables. In our case, the phenotype often has a baseline level that is not zero and may have variability due to factors unrelated to gene expression (e.g., environmental factors, baseline biological activity). This baseline level would be captured

by an intercept. This suggests that without an intercept, the model may incorrectly assume that the phenotype is zero when all gene expression levels are zero, which may not be biologically meaningful. Excluding the intercept forces the regression line to pass through the origin, potentially introducing bias into the model if the true relationship doesn't naturally do so. For gene expression and phenotype, this is unlikely, as some baseline level of phenotype exists even in the absence of gene expression contributions. Hence, an intercept should be included in the regression model. This will allow the model to account for the baseline level of the phenotype when no gene expression effects are present.

4.3 Dimensionality Reduction

Dimensionality reduction is an important preprocessing step in data analysis. It addresses several key objectives:

- Reduces computational complexity.
- Mitigates the curse of dimensionality.
- Improves model performance.

PCA is one of the most commonly used dimensionality reduction methods. However, PCA itself doesn't identify specific significant regressors; it identifies patterns or combinations of regressors. Considering the objective of the study, correlation-based dimensionality reduction is preferred. In the following method, the correlation value between each of the regressors (i.e., genes) and the targeted variable (i.e., phenotype) is calculated. The predictors are then ranked by their absolute correlation values. The top 200 predictors with the highest correlation are selected.

4.4 High-Dimensional Regression Approach

Even after performing dimensionality reduction, the data still consisted of a large number of variables ($p = 200$) and a small sample size ($n = 50$). Since $p \gg n$, the data is still high-dimensional. To perform regression on high-dimensional data, certain assumptions need to be verified. These assumptions are:

- Multicollinearity.
- Normality of residuals.
- Linearity.
- Homoscedasticity.
- Independence of residuals.

4.5 Ridge and Lasso Regression

Lasso Regression: Adds a penalty proportional to the absolute value of the coefficients. It performs both regularization and feature selection, as some coefficients are shrunk to exactly zero, excluding irrelevant features.

$$\text{Loss} = \text{RSS} + \lambda \sum_{j=1}^{p} \beta_j \tag{2}$$

Ridge Regression: Shrinks coefficients by adding a penalty proportional to the square of the coefficients to the loss function. Helps in managing multicollinearity but does not perform feature selection, as coefficients are only reduced, not set to zero.

$$\text{Loss} = \text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{3}$$

4.6 Bayesian Regression

Bayesian regression treats model parameters (β) as random variables with prior distributions. It combines prior information with observed data to estimate a posterior distribution, offering a probabilistic framework that captures uncertainty in predictions.

5 Results

Model	**MSE**	**R-Squared**	**Alpha**
Ridge Regression	8646.7006	0.9302	0.01
Lasso Regression	5250.1754	0.9576	0.01
Bayesian Regression	8645.9589	0.9302	—

Table 1: Performance of Regression Models

Key Observations:

- The MSE value for lasso is less as compared to other methods and also R squared for lasso is highest, suggesting that lasso regression is the best fit for the model.
- The reason why Lasso works better is because it shrinks coefficients of less relevant regressors to exactly zero, whereas ridge does not perform feature selection.
- As lasso is used for feature selection, the model retained 11 features, i.e., out of the 200 selected features, only 11 genes played a significant role in determining the phenotype.

Gene 1	**Gene 2**	**Gene 3**	**Gene 4**	**Gene 5**	**Gene 6**	**Gene 7**	**Gene 8**	**Gene 9**	**Gene 10**	**Gene 11**
X_2	X_1	X_{430}	X_{273}	X_{596}	X_{1824}	X_{671}	X_{1562}	X_4	X_3	X_5

Table 2: Significant Genes Identified by Lasso Regression

- From the study, it was found that Bayesian regression is not a better fit than the lasso model.

6 Conclusion

The study demonstrates that:

- The MSE value for lasso is less as compared to other methods and also R squared for lasso is highest, suggesting that lasso regression is the best fit for the model.
- The reason why Lasso works better is because it shrinks coefficients of less relevant regressors to exactly zero, whereas ridge does not perform feature selection.
- As lasso is used for feature selection, the model retained 11 features, i.e., out of the 200 selected features, only 11 genes played a significant role in determining the phenotype.
- These 11 genes are: $X2, X1, X430, X273, X596, X1824, X671, X1562, X4, X3, X5$.
- From the study, it was found that Bayesian regression is not a better fit than the lasso model.