Question

The goal of this simulation study is to evaluate the predictive performance of different regression models—Linear Regression, Ridge Regression, and Lasso Regression—when estimating an outcome based on exoplanet parameters. We aim to determine whether the discovery method (Transit vs. Non-Transit) influences model performance due to differences in the underlying distribution of predictors.

Data

In this simulation, we create synthetic data that mimics the key features of exoplanets, based on three main components: stellar properties, planetary properties, and orbital characteristics. These three variables are generated from a multivariate normal distribution, meaning they are normally distributed and moderately correlated with each other (correlation = 0.6).

We then use a linear equation to generate the response variable yyy, which depends on the three predictors. The formula is:

$y = 0 \cdot X1 + 2.5 \cdot X2 - 3 \cdot X3 + \text{random noise}$

The random noise is normally distributed with a mean of 0 and a standard deviation of 1. This setup reflects a situation where only two of the predictors actually influence the outcome, helping us test how well different models can detect the true relationships.

Estimates

We estimate Mean Squared Error (MSE) as the performance metric for each model. The simulation calculates MSE separately for each of the three models and two discovery methods.

Methods

We evaluate the following modeling approaches: Linear Regression, Ridge Regression, and Lasso Regression. We hypothesize that Ridge and Lasso will outperform Linear Regression due to the presence of multicollinearity among predictors.

Performance Criteria

The simulation uses the following criteria to assess model performance: mean MSE across simulations for each model-method combination, spread of MSE (using boxplots) to capture model variance, and comparisons between Transit and Non-Transit groups to assess

group-specific differences. These metrics summarize performance across all repetitions, giving insight into both accuracy and reliability.

Simulation Plan

Number of Simulations: 1000
Predictor Correlation: rho = 0.6
Noise Level: sigma = 1

Challenges or Limitations

Model assumptions: All methods assume a linear relationship between predictors and outcome. Real-world data may violate this.
Random variability: Sampling-based randomness (in noise, discovery method assignment, and train-test split) may introduce variability in MSE.