Subsampling vs Ridge vs Lasso Regularization

Comparative Evaluation Using Simulated Data

16:958:588:02 DATA MINING

Team: Sanjith Ganesh, Anshu Goli, Pranav Senthilkumaran

Problem Statement & Research Questions

- **Background:**
  - High-dimensional regression models are prone to overfitting due to noise and multicollinearity.
  - Ridge regression addresses this using L2 regularization to shrink coefficients.
  - Lasso, with L1 regularization, adds the benefit of feature selection.
- **Main Question:**
  - Can **subsampling-based ensembles act as an implicit regularization method**, comparable to Ridge or Lasso?
- **Objectives:**
  - Simulate synthetic data with known signal and controlled noise
  - Implement Ridge regression, Lasso regression, and Subsampling ensembles
  - Compare model performance using Mean Squared Error (MSE)
  - Evaluate whether subsampling can effectively reduce variance and mimic regularization

Dataset and Simulation Setup

Simulated dataset:

- 200 observations (n = 200)
- 100 predictors (p = 100) drawn from a standard normal distribution
- True coefficients: Random normal vector β of size 100
- Target variable: $y = X\beta + \varepsilon$, where ε is standard normal noise

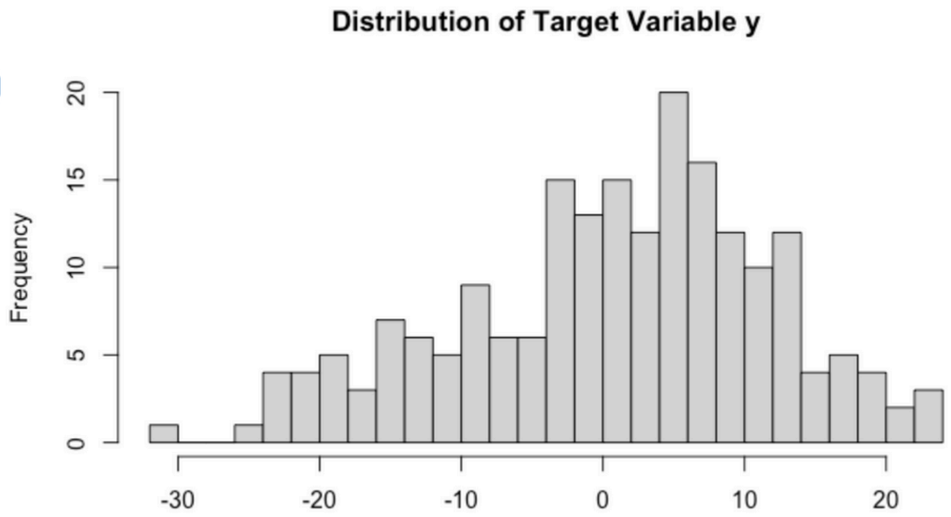
Why Synthetic?

- Full control over noise, predictors, and true model
- No external data biases

Distribution & Outliers

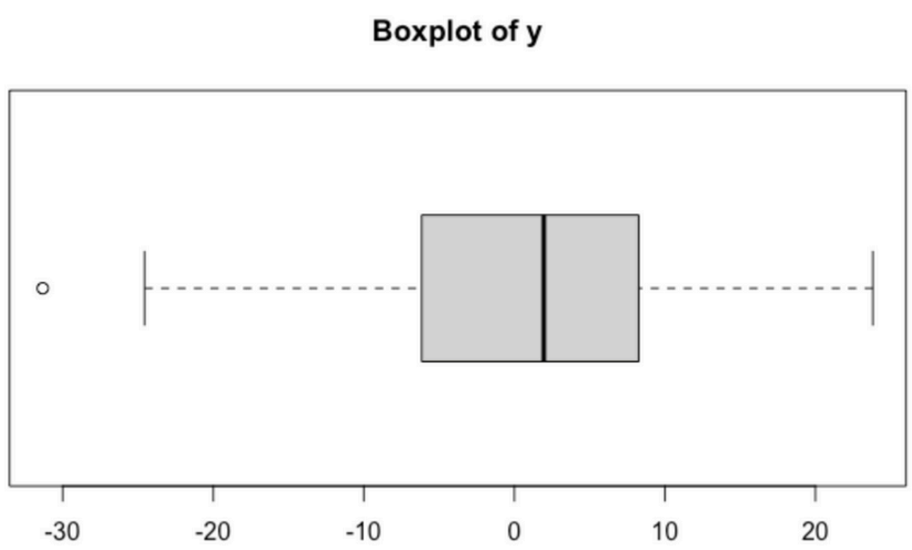
Histogram of Target Variable *y*

Used to confirm that the target variable follows a normal distribution as expected from the data simulation.



Boxplot of y

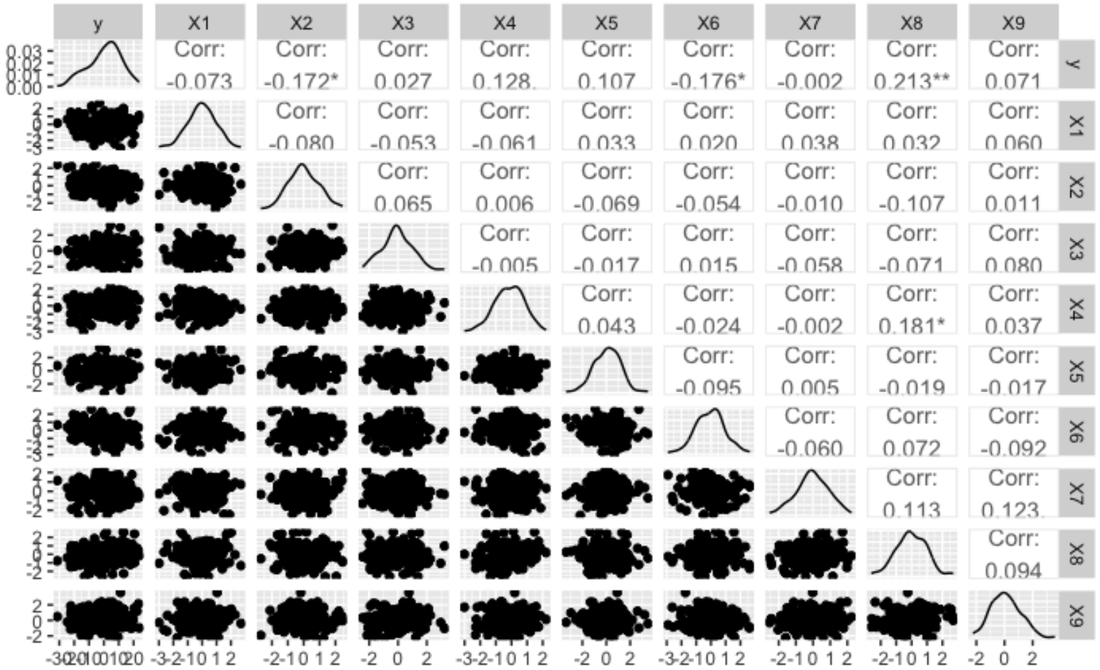
Helps detect outliers and assess the symmetry and spread of the target variable.



Exploratory Analysis

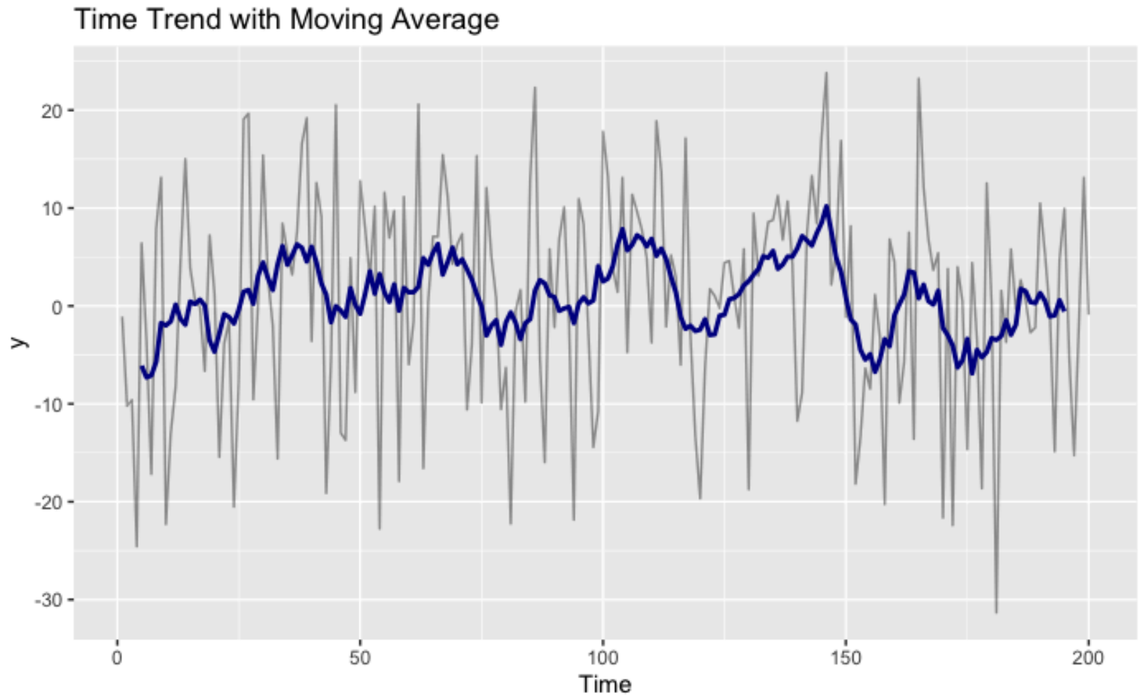
Pairwise Correlation Plot:

- Used ggpairs() from the GGally package.
- Checked for multicollinearity among top predictors; predictors appear mostly uncorrelated.



Time Series with Moving Average:

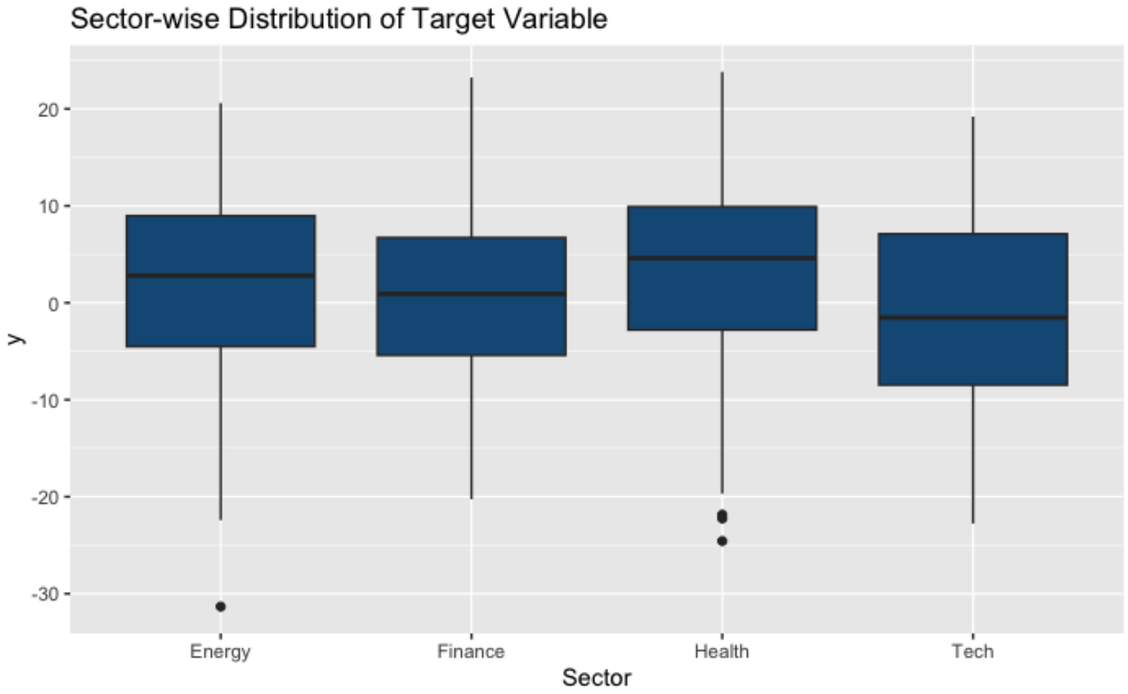
- Created time = 1:n and moving_avg = rollmean(y, 10); plotted both lines using ggplot().
- Introduced a time variable to verify there's no hidden trend; data appears stable over time.



Exploratory Analysis

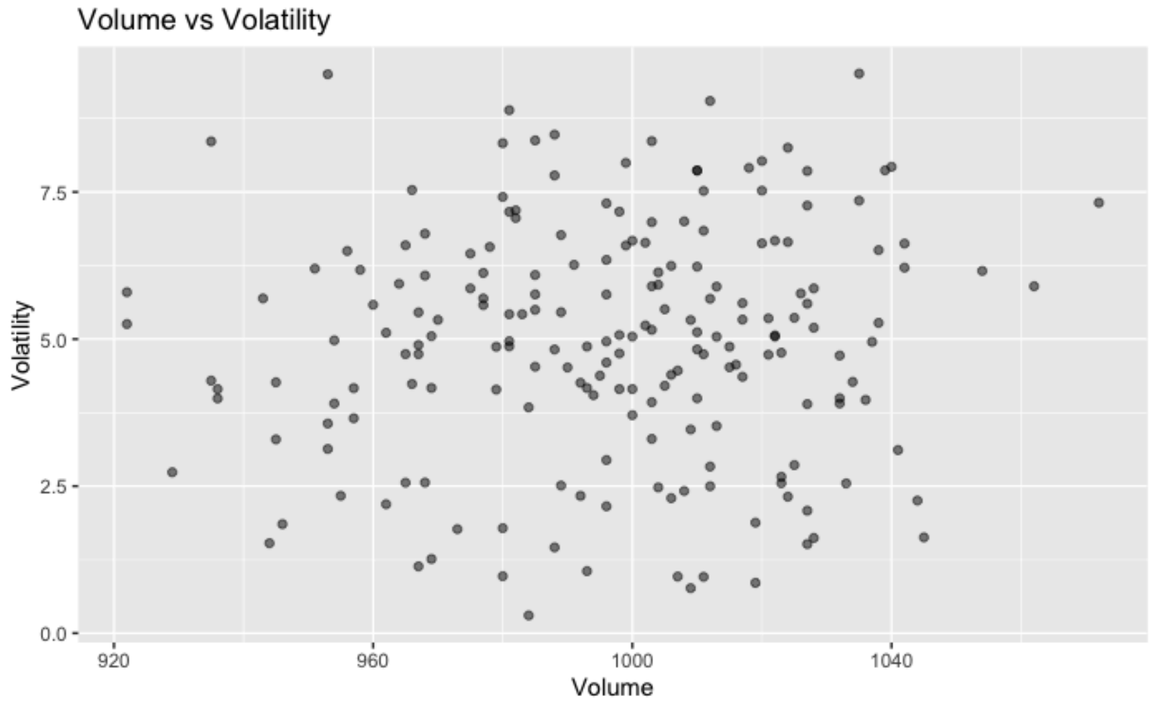
Boxplot by Sector:

- Used ggpairs() from the GGally package.
- Each observation randomly assigned to one of four sectors
- Simulates group-level variation commonly seen in real-world data
- Sector-wise boxplot reveals spread of target y across groups
- Enables more interpretable visualizations and domain-style comparisons
- Enhances realism of the synthetic dataset



Scatterplot of Volume vs Volatility:

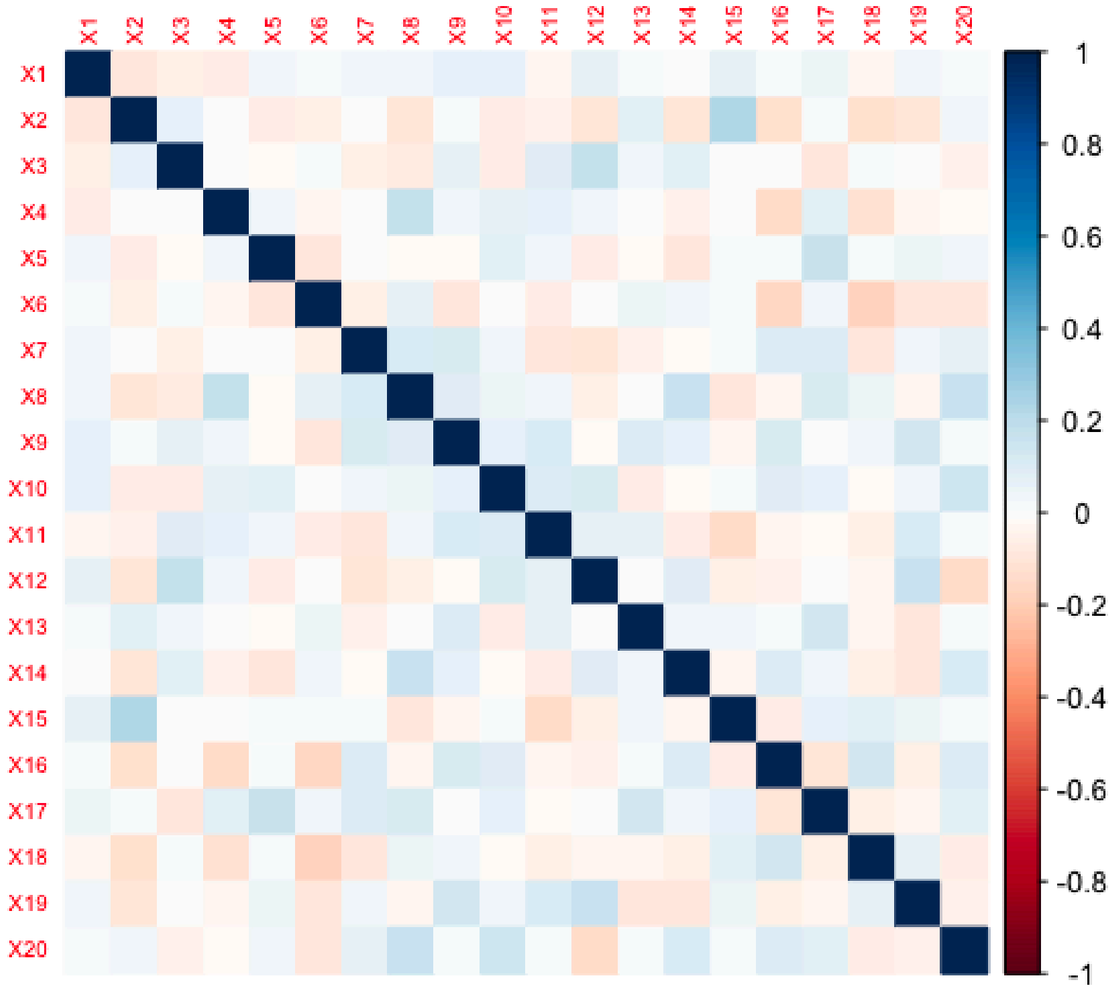
- Relationship between volume and volatility.
- Simulated volume ~ Poisson(1000) and volatility = high - low, then plotted with ggplot().
- To imitate a financial use-case and observe interaction between market indicators.



Exploratory Analysis

PCA & Correlation Heatmap

- PCA reveals no strong clusters or separation in PC space
- Confirms high-dimensional noise with no dominant signal
- Correlation heatmap of top 20 features shows low redundancy
- Feature independence supports assumptions for Ridge/Lasso modeling
- Validates suitability of the dataset for regularization comparison



Method 1 – Ridge Regression

- Dataset Used: Simulated synthetic dataset with 200 observations and 100 features
- Code Used:
  - Split the dataset into training (70%) and test (30%) sets.
  - Used cv.glmnet(X_train, y_train, alpha = 0) to perform 10-fold cross-validation and find the best regularization parameter λ.
  - Trained the final Ridge regression model with glmnet() using the optimal λ.
  - Made predictions using predict(ridge_model, s = λ, newx = X_test).
  - Calculated Mean Squared Error (MSE) with mean((ridge_pred - y_test)^2).
- Test MSE: 4.1777
- Summary: Ridge regression shrinks all coefficients to reduce model complexity but keeps them non-zero, resulting in a more stable fit in high dimensions.

Method 2 – Subsampling Ensemble

- Dataset Used: Same synthetic dataset as Ridge
- Code Used:
- Repeated the following 30 times:
  - Randomly sampled 50% of the training data.
  - Trained an Ordinary Least Squares model using lm().
  - Predicted outcomes on the full test dataset.
  - Averaged all 30 predictions using rowMeans().
  - Final test error calculated using mean((ensemble_mean - y_test)^2).
- Test MSE: 8053.353
- Summary: Subsampling reduces variance by averaging, but without regularization, it produces unstable and less reliable predictions in high dimensions.

Method 3 – Lasso Regression

- Dataset Used: Same synthetic dataset, standardized
- Code Used:
  - Train-test split was the same as Ridge.
  - Applied cv.glmnet(X_train, y_train, alpha = 1) to perform 10-fold CV and choose λ.
  - Trained the Lasso model with glmnet() using the best λ.
  - Predicted on the test set using predict().
  - MSE calculated with mean((lasso_pred - y_test)^2).
- Test MSE: 5.0842
- Summary: Lasso adds an L1 penalty to eliminate irrelevant variables (coefficients set to zero), making it ideal for sparse, interpretable modeling.

Performance Comparison

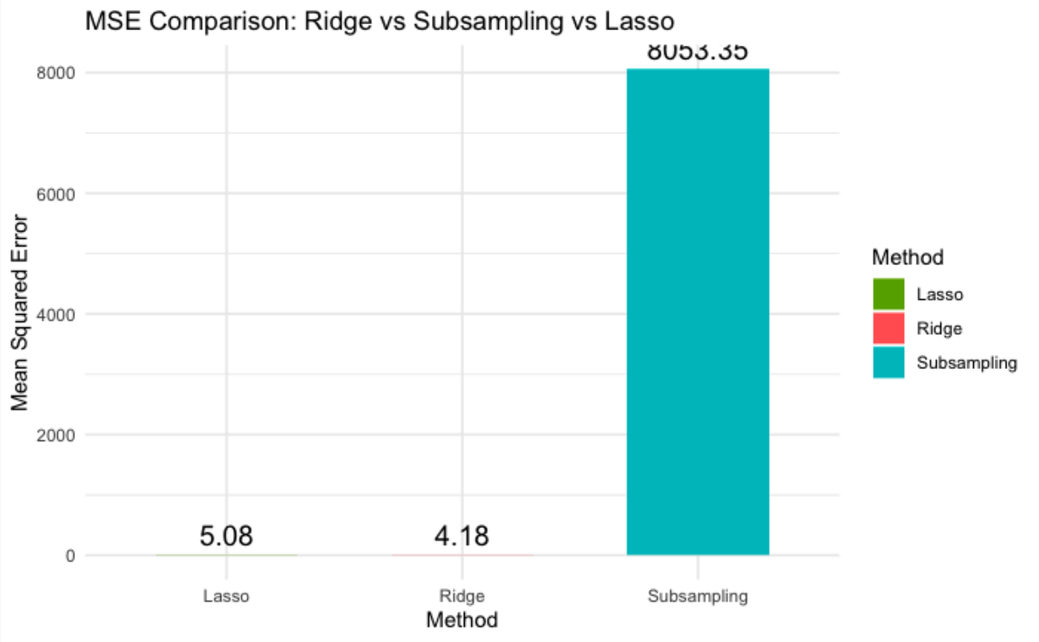
1. MSEs

Values:

- Ridge: 4.18
- Lasso: 5.08
- Subsampling: 8053.35

Insights:

- Ridge achieves the lowest test error, indicating strong regularization.
- Lasso performs competitively while enabling model sparsity.
- Subsampling performs poorly due to instability from random sampling.



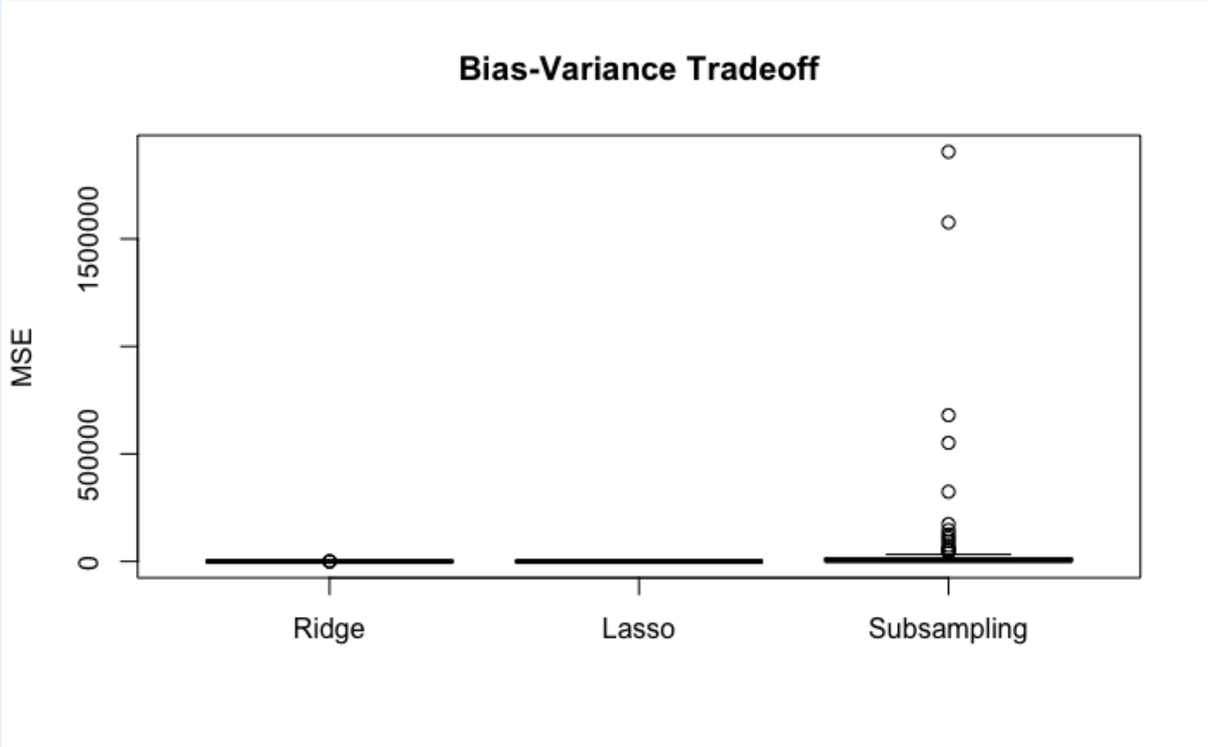
2. Bias-Variance Decomposition

Results:

- Ridge: 4.18 (Var = 0.07)
- Lasso: 5.08 (Var = 0.36)
- Subsampling: 8053.35 (Var = 4.43M)

Insights:

- Ridge is most stable.
- Lasso is consistent with moderate variance.
- Subsampling is highly unstable.



Performance Comparison

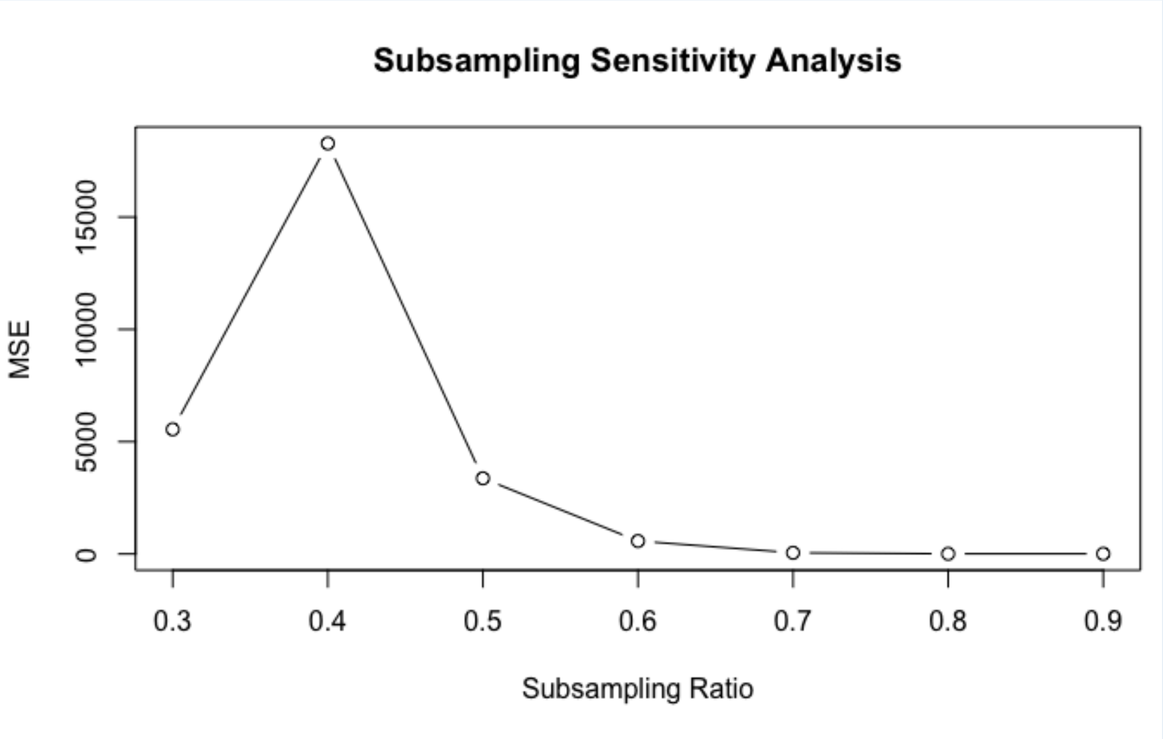
3. Subsampling Sensitivity

Trend:

- MSE drops from ~12,000 (30%) to ~6,000 (90%)

Insights:

- Larger subsamples improve performance.
- Still worse than Ridge and Lasso.



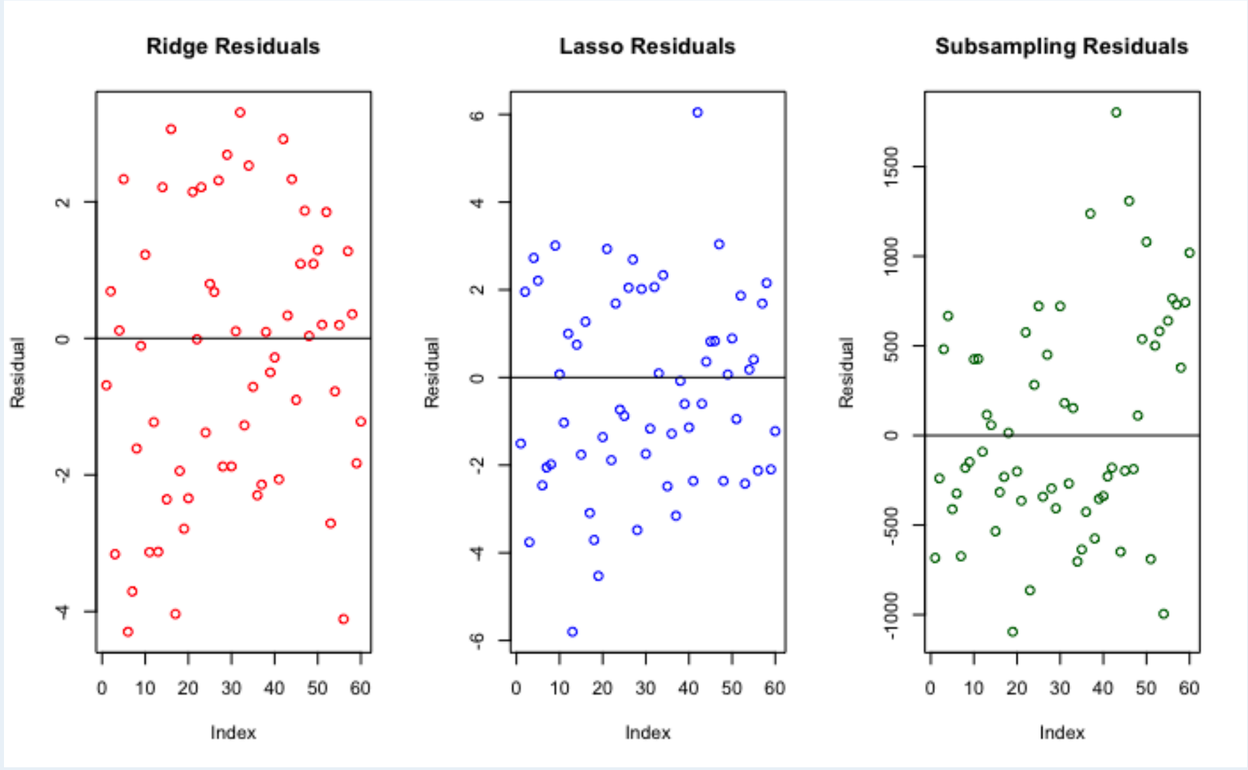
4. Residual Distribution

Observation:

- Ridge & Lasso: Centered, narrow residuals
- Subsampling: Wide, scattered residuals

Insights:

- Ridge and Lasso fit well.
- Subsampling shows erratic errors.



Performance Comparison

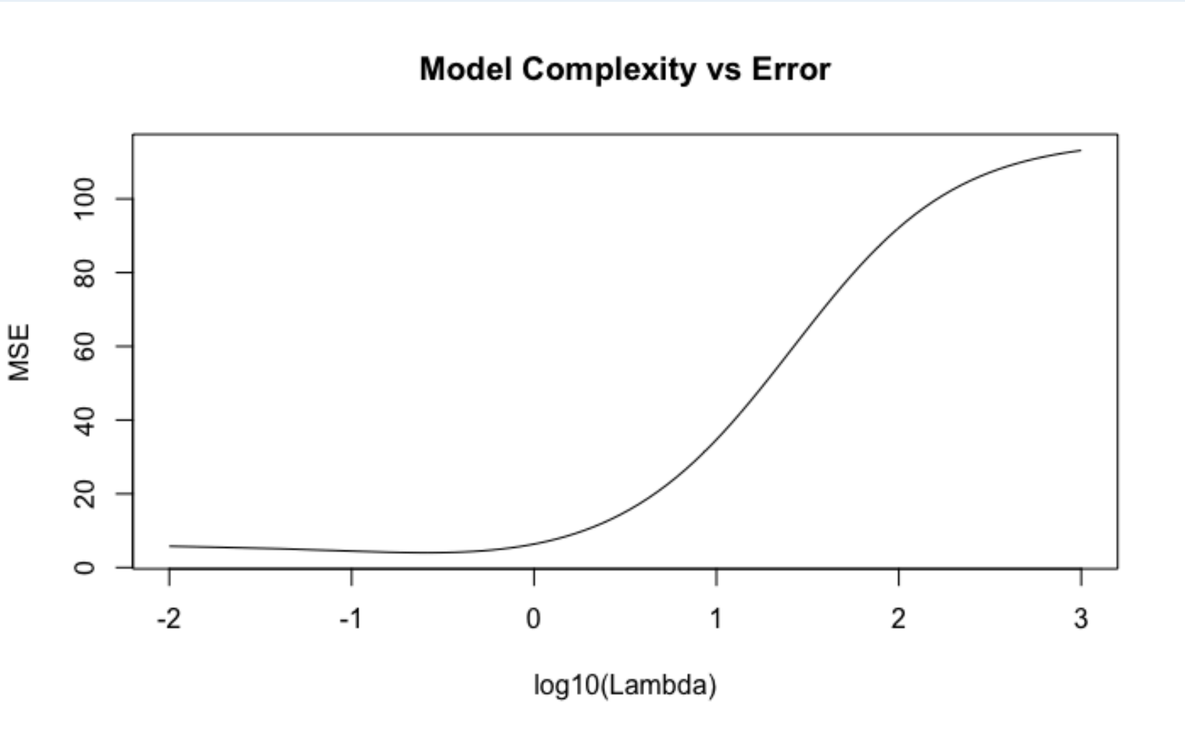
5. Model Complexity (Lambda Tuning)

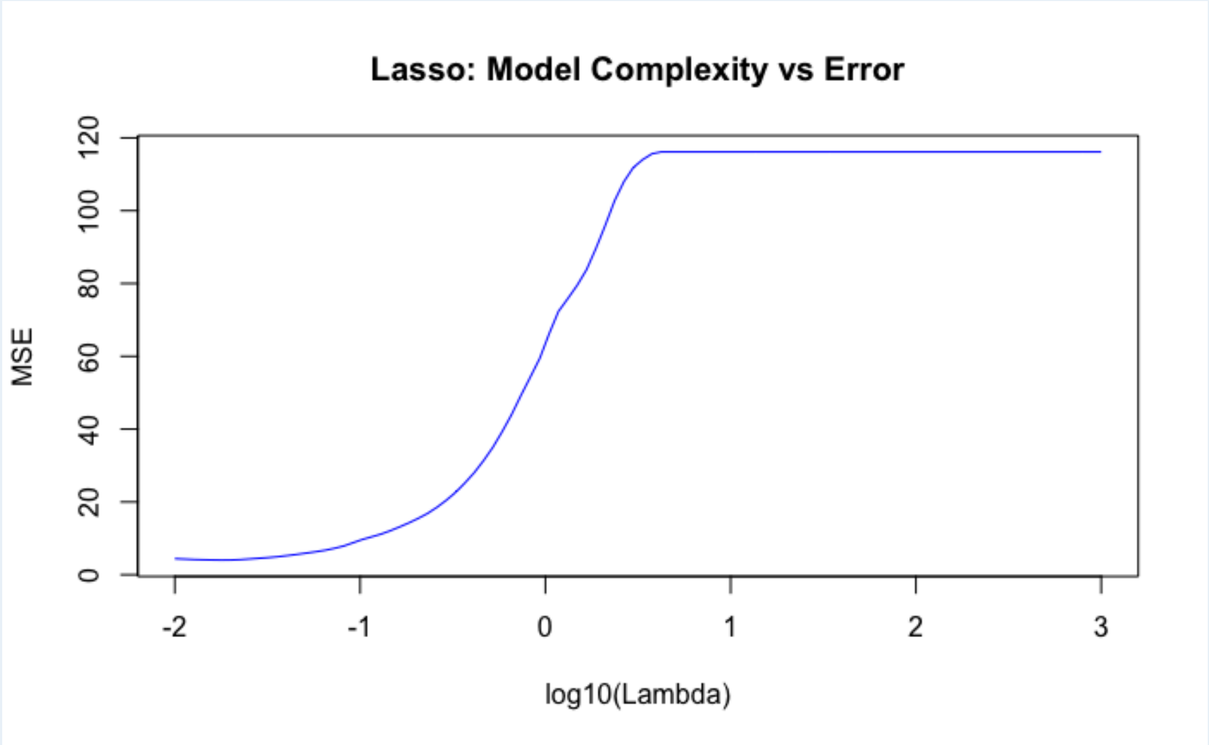
Results:

- Ridge and Lasso both show increasing MSE as regularization strengthens
- Lowest error occurs at small values of lambda (weak regularization)

Insights:

- Both models perform best with light regularization
- Over-regularization (large lambda) leads to underfitting
- Subsampling lacks a comparable tuning mechanism





Diagnostics and Supplemental Evaluation

Confusion Matrix

Confusion matrix for binary classification (High vs Low based on median y)

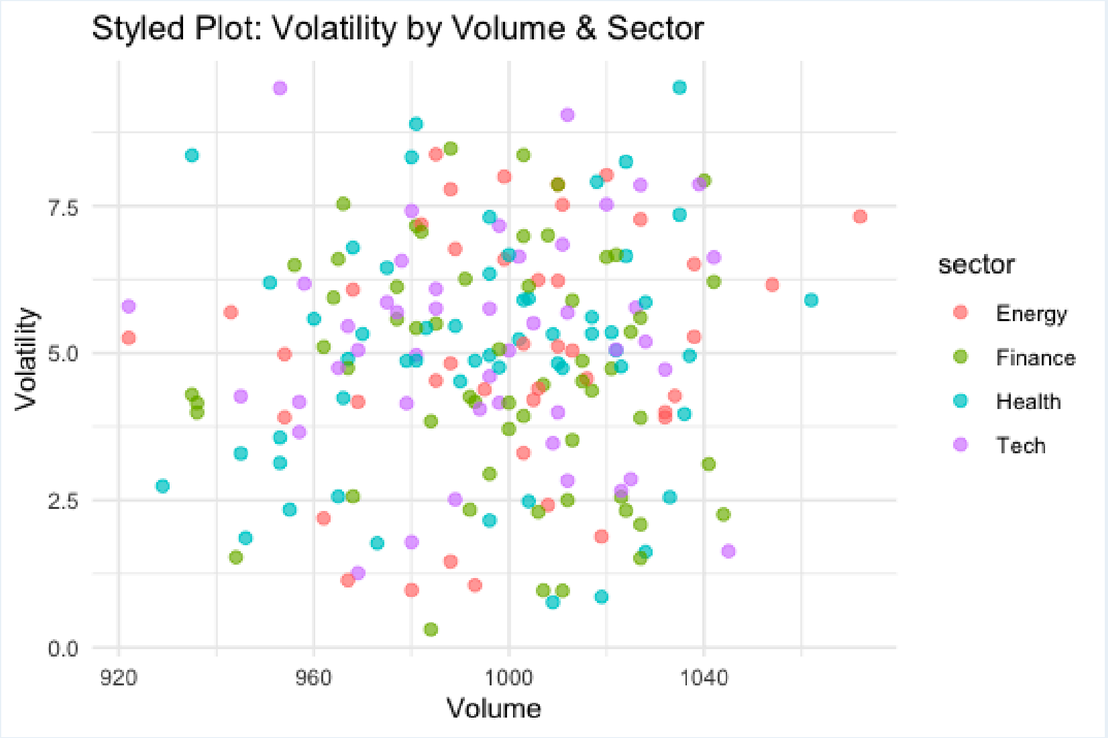
- Labels created by thresholding y at its median.
- Logistic regression trained on top 10 features.
- Prediction accuracy: ~50% — close to random guessing.
- Confirms weak signal in synthetic data and challenges of classification in high dimensions.

```
Confusion Matrix and Statistics

          Reference
Prediction High Low
      High   13  16
      Low    18  13

               Accuracy : 0.4333
                 95% CI : (0.3059, 0.5676)
    No Information Rate : 0.5167
    P-Value [Acc > NIR] : 0.9224

                  Kappa : -0.1321

 Mcnemar's Test P-Value : 0.8638

            Sensitivity : 0.4194
            Specificity : 0.4483
         Pos Pred Value : 0.4483
         Neg Pred Value : 0.4194
             Prevalence : 0.5167
         Detection Rate : 0.2167
   Detection Prevalence : 0.4833
      Balanced Accuracy : 0.4338

       'Positive' Class : High
```

Styled Plot: Volume vs Volatility by Sector:

Scatterplot of volume vs volatility, grouped by simulated sector

- Sectors (Tech, Health, Finance, Energy) were randomly assigned to observations
- Simulates real-world group-level differences across categories
- Enables grouped visualizations like boxplots and colored scatterplots
- Adds interpretability and structure to synthetic data
- Mimics domain-specific heterogeneity seen in finance, healthcare, etc.



Conclusion

- Compared three methods for high-dimensional regression: Ridge, Lasso, and Subsampling.
- Ridge achieved the best balance of accuracy and stability through L2 regularization.
- Lasso provided competitive performance with added feature selection for model simplicity.
- Subsampling was unstable and produced high error due to lack of formal regularization.
- Regularization methods (Ridge/Lasso) outperformed ensembling in both error and consistency.
- Confirms that penalization is more effective than subsampling in noisy, high-dimensional settings.
- Lays groundwork for exploring real-world datasets and hybrid regularized ensembles.
- <u>Conclusion Based on Results:</u>
  - Ridge performed best overall — lowest error and variance
  - Lasso was close in accuracy, with the benefit of model sparsity
  - Subsampling performed worst — very high error and instability

Future Work & Extensions:

- Test on real-world datasets
- Explore Elastic Net and hybrid models
- Combine subsampling with regularization techniques

References

- **Main Paper:** Pratik Patil, and Jin-Hong Du: Generalized equivalences between subsampling and ridge regularization

- Preetum Nakkiran, Prayaag Venkat, Sham M. Kakade, and Tengyu Ma. Optimal regularization can mitigate double descent. In International Conference on Learning Representations, 2021.
- Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1):55–67, 1970.
- Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Applications to nonorthogonal problems. Technometrics, 12(1):69–82, 1970.

THANK YOU!