Multiple Linear Regression Analysis Using R

Jishnu Sai Matra

This paper compares ordinary least squares (OLS) regression with Lasso regression using the Boston Housing dataset. While both models demonstrated similar predictive performance (R^2 0.74), they differ in their treatment of predictor variables. Lasso regression incorporates L1 regularization, which shrinks coefficient estimates and can perform automatic feature selection. The analysis illustrates practical implementation of both methods and discusses their respective advantages and limitations. Results suggest that while regularization had minimal impact on model fit for this dataset, Lasso regression remains valuable for feature selection and improving model interpretability.

Introduction

Regression analysis is fundamental in statistical modeling, allowing researchers to understand relationships between variables and make predictions. While ordinary least squares (OLS) regression remains widely used, it can suffer from overfitting when dealing with multiple predictors, particularly in datasets where predictors are correlated. Lasso (Least Absolute Shrinkage and Selection Operator) regression offers an alternative approach by incorporating regularization, which constrains coefficient estimates and can perform automatic feature selection.

This handout demonstrates the practical implementation and comparison of linear regression and Lasso regression using the Boston Housing dataset. The analysis aims to illustrate how regularization affects model coefficients and predictive performance, providing insights into when Lasso regression may be preferable to traditional linear regression.

Methods

Dataset

The Boston Housing dataset, available in the MASS package in R, contains information on housing values in suburbs of Boston. The dataset includes 506 observations across 14 variables.

The outcome variable is `medv` (median value of owner-occupied homes in $1000s), while the 13 predictor variables include per capita crime rate, proportion of residential land zoned for lots, average number of rooms per dwelling, and other neighborhood characteristics.

Statistical Approach

Two regression models were fitted to the data. First, an ordinary least squares regression model was estimated using the `lm()` function in R, which provides a baseline without regularization. Second, a Lasso regression model was fitted using the `glmnet` package, which adds an L1 penalty term to the loss function. The optimal regularization parameter ( ) was selected through 10-fold cross-validation (Harris (2023)).

The Lasso regression objective function minimizes:

$$\text{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|$$

where RSS is the residual sum of squares, is the regularization parameter, and β_j represents the regression coefficients. This penalty encourages sparsity by shrinking some coefficients to exactly zero. Model performance was evaluated using R^2, which measures the proportion of variance in the outcome variable explained by the predictors.

Implementation

The analysis was conducted in R (version 4.5.1). First, the necessary packages were loaded and the data prepared:

```
# Load required packages
library(glmnet)
library(MASS)

# Load and prepare data
data(Boston)
X <- as.matrix(Boston[, -14])  # Predictor matrix
y <- Boston$medv                # Outcome variable
```

Linear Regression Model

The baseline OLS model was fitted using all available predictors:

```
# Fit linear model
lm_model <- lm(medv ~ ., data = Boston)
lm_coef <- coef(lm_model)
lm_r_squared <- summary(lm_model)$r.squared
```

The linear model yielded an R^2 of 0.7406, indicating that approximately 74.1% of the variance in housing values is explained by the predictors.

Lasso Regression Model

For the Lasso model, cross-validation was used to determine the optimal value:

```
# Fit Lasso model with cross-validation
cv_model <- cv.glmnet(X, y, alpha = 1)
best_lambda <- cv_model$lambda.min

# Fit final model with optimal lambda
lasso_model <- glmnet(X, y, alpha = 1, lambda = best_lambda)
lasso_coef <- coef(lasso_model)

# Calculate predictions and R-squared
lasso_predictions <- predict(lasso_model, newx = X)
lasso_r_squared <- 1 - (sum((y - lasso_predictions)^2) /
                        sum((y - mean(y))^2))
```

The cross-validation procedure selected $= 0.0193$. The resulting Lasso model achieved an R^2 of 0.7404.

Results

Model Performance

Both models demonstrated comparable predictive performance on the training data. The OLS regression achieved an R^2 of 0.7406, while the Lasso regression achieved an R^2 of 0.7404. The difference in R^2 values was minimal (3×10^{-4}), indicating that the regularization penalty had

little impact on in-sample fit. This similarity in performance suggests that while Lasso provides coefficient shrinkage, it does not substantially sacrifice explanatory power in this dataset.

Discussion

This analysis highlights Lasso regression as a practical alternative to traditional OLS regression. Both methods produced similar R^2 values, but differ in handling predictors: OLS retains all predictors with unconstrained coefficients, whereas Lasso applies regularization, shrinking coefficients and enabling automatic feature selection.

The similar performance indicates that the Boston Housing dataset does not suffer from severe multicollinearity or overfitting. In datasets with more predictors or stronger correlations, Lasso's benefits would be more evident. Note that the reported R^2 values are in-sample estimates; testing on held-out data would better assess generalization.

Lasso offers advantages in interpretability by identifying influential predictors and controlling large coefficients, improving model stability. However, it assumes sparsity in true coefficients, and highly correlated predictors may lead Lasso to arbitrarily select variables. In such cases, methods like elastic net, which combines L1 and L2 penalties, may be preferable.

Conclusion

This handout illustrated the implementation and comparison of linear and Lasso regression using the Boston Housing dataset. Both models achieved similar predictive performance, with R^2 values above 0.73. The Lasso model successfully applied regularization to shrink coefficient estimates while maintaining explanatory power. These results highlight that Lasso regression can be a valuable tool in the statistical modeling toolkit, particularly when feature selection or improved generalization is desired. Future analyses could extend this comparison by evaluating model performance on independent test data and exploring alternative regularization approaches.

Affidavit

I hereby affirm that this submitted paper was authored unaided and solely by me. Additionally, no other sources than those in the reference list were used. Parts of this paper, including tables and figures, that have been taken either verbatim or analogously from other works have in each case been properly cited with regard to their origin and authorship. This paper either in parts or in its entirety, be it in the same or similar form, has not been submitted to any other examination board and has not been published.

I acknowledge that the university may use plagiarism detection software to check my thesis. I agree to cooperate with any investigation of suspected plagiarism and to provide any additional information or evidence requested by the university.

Checklist

- ☒ The handout contains 7 pages of text.
- ☒ The submission contains the Quarto file of the handout.
- ☒ The submission contains the Quarto file of the presentation.
- ☒ The submission contains the HTML file of the handout.
- ☒ The submission contains the HTML file of the presentation.
- ☒ The submission contains the PDF file of the handout.
- ☒ The submission contains the PDF file of the presentation.
- ☒ The title page of the presentation and the handout contain personal details (name, email, matriculation number).
- ☒ The handout contains a bibliography, created using BibTeX with an APA citation style.
- ☒ Either the handout or the presentation contains R code that demonstrates coding expertise.
- ☒ The filled out Affidavit.
- ☒ The link to the presentation and the handout published on GitHub.

[Jishnu Sai Matra,] [10 Dec 2025,] [Leverkusen]

Harris, M. (2023). *Regression with cross-validation in r.* https://www.statswithr.com/tutorials/regression-with-cross-validation-in-r