PS9_RYU

April 2024

1 Problems

1.1 Q7.

The dimension of housing_train: observations=404 & variables=14

The dimension of housing_train_preppaed: observations=404 & variables=75

How many more X variables do you have than in the original housing data?
: more 61 variables than original housing data

1.2 Q8.(Lasso)

In-sample RMSE was 0.413
Out-of-sample RMSE is 0.390
In-sample $\hat{R2}$ was 0
Out-of-sample $\hat{R2}$ is -0.000467
rsme: 0.170
rsq: 0.809
penalty: 0.00139
mean: 0.0632
n : 6
std_err: 0.00503

1.3 Q9.(Ridge)

In-sample RMSE was 0.197
Out-of-sample RMSE is 0.194
In-sample $\hat{R2}$ was 0.773
Out-of-sample $\hat{R2}$ is 0.752
rsme: 0.173
rsq: 0.802
penalty: 0.0373
mean: 0.0716

n : 6
std_err: 0.00585

1.4 Q10.

1) Would you be able to estimate a simple linear regression model on a data set that had more columns than rows?

: In a simple linear regression model, if a dataset has more columns (variables) than rows (observations), we can refer it as high-dimensional. In this case, estimating a simple linear regression model is possible, but there are potential issues that need to be approached carefully.

First of all, We can think of degree of freedom issue. With more columns than rows, the degrees of freedom in estimating the model parameters will be limited, leading to overfitting problem. Overfitting problem can negatively impacts when generalizing to unseen data. And then, we can think of multicollinearity issue, which occurs when the independent variables are highly correlated with each other. This can lead to a biased estimate of the regression coefficient.

2) comment on where your model stands in terms of the bias-variance trade-off.

The goal is to find the balance between bias and variance that minimizes the total error. But, we can find that in general, decreasing one will often increase the other, called the bias-variance tradeoff. RMSE combines both the bias and the variance of the model. In the previous two models, RMSE of Lasso is 0.170 and RMSE of Ridge is 0.173, indicating that estimate of Lasso can be a more reasonable and proper model compared to Ridge.