Problem Set 9 - ECON 5253

Thomas Mondry

April 11, 2022

1 Results

The results were obtained through the following process, as in the lecture 20 example code:

1. Subset the UCI housing data into training (80%) and test (20%)
2. Prepare both datasets using the recipe in the description
3. Tune the penalty hyperparameter λ with a fixed L1/L2 mixture hyperparameter (i.e., 1 for LASSO and 0 for ridge)
4. Estimate the respective model with the fitted λ
5. Evaluate the model on both the full training set (in-sample) and the test set (out-of-sample), using RMSE as the metric.

The results of this process are displayed in the table:

model	fitted λ	In-sample RMSE	Out-of-sample RMSE
LASSO	0.00139	0.13652	0.18789
ridge	0.03728	0.19960	0.18437

2 Commentary

The data consists of 506 observations, of which 404 were selected into the training set and 102 into the test set. The initial raw data contained 13 features, while the prepared data has 74.

One less than the number of observations, $n - 1$, is a hard upper limit for the number of columns in the OLS regression model because OLS is a deterministic method to find a singular solution, and there is no singular solution when $k > n - 1$.

We can reasonably expect that when we tune hyperparameters appropriately, we will estimate a model that is near the error-minimizing level of the bias-variance tradeoff. This data is not enormous, so even though they were selected randomly, there is the potential for one or more of the CV folds or for the test set to differ from the others in a way that affects model performance metrics; however, using CV is the best we can do. Looking at the RMSE metrics, we are seeing similar performance in the training set and the test set, which performs slightly better in one model and slightly worse in another. This suggests that we are neither significantly overfitting nor significantly underfitting.