

Lab 3

Statistical learning - EEBL

2023-10-12

Model selection

- Randomly split the dataset into training and test sample so to retain 60% of observations for training.
- Consider the complex model:

```
model_full = price ~ sqft + I(sqft ^ 2) + I(sqft ^ 3) + Age + I(Age ^ 2) +  
  I(Age ^ 3) + I(sqft * Age) + I(sqft ^ 2 * Age) + I(sqft * Age ^ 2) +  
  Pool + Baths + I(Baths ^ 2) + I(Baths ^ 3) + Bedrooms + I(Bedrooms ^ 2) +  
  I(Bedrooms ^ 3) + Fireplace + Waterfront + DOM + factor(Occupancy) + factor(Style)
```

how many regressors does it include?

- Perform forward subset selection while making sure that the variable **sqft** is always included in the regression. How many regressors does the model with the lowest BIC have?
- Consider the model attaining the lowest Cp:
 - (i) which variables does it include?
 - (ii) does it perform better than the full model in predicting house prices out of sample?

Principal component regression

- Consider the eigendecomposition of the correlation matrix. What do eigenvalues and eigenvectors represent?
- Consider the scatterplot of the first four principal components. Do they exhibit any linear relationship?
- Perform PCR: can we improve out of sample by fitting a reduced model selected according to AIC?