Problem Set - Model Selection

Code

Rich Sweeney & Charlie Murry

Harrison & Rubinfeld (JEEM, 1978 (https://www-sciencedirect-com.proxy.bc.edu/science/article/pii/0095069678900062)) conducted one of the first hedonic property value studies of willingness to pay for clean air. Their sample consisted of 506 Census tracts in Massachussets. Their primary interest was the relationship between NOx (https://www3.epa.gov/region1/airquality/nox.html) concentration levels and median home values.

A slightly modified version of the HR data (which has been used extensively due to it's inclusion in R's `MASS` package), has been provided in the file `boston_cl.csv`

There are 506 observations, representing Boston census tracts.

Variables in order:
- CRIM: per capita crime rate by town
- ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS: proportion of non-retail business acres per town
- CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise) - NOX: nitric oxides concentration (parts per 10 million)
- RM: average number of rooms per dwelling
- AGE: proportion of owner-occupied units built prior to 1940
- DIS: weighted distances to five Boston employment centres
- RAD: index of accessibility to radial highways
- TAX: full-value property-tax rate per $10,000
- PTRATIO: pupil-teacher ratio by town
- BLACK: $1000(Bk - 0.63)^2$ where Bk is the proportion of African-Americans by town
- LSTAT: % lower status of the population
- MEDV: Median value of owner-occupied homes in $1000's (topcoded at 50K)

Code

Set aside 20 percent of the data as a test sample. For each of the models specified below, estimate them using only the remaining 80 percent of the data (training data).

Code

1. How correllated are these variables?

Code

2. Estimate the original HR model using the training data.

Project the the log(Median House Price) onto all of the other variables. Everything should enter linearly, except for `NOx` and `RM`, which should only enter quadratically.

Code

3. Now estimate the model using LASSO.

Use k=10 fold cross validation to select lambda. Select the model with the largest lambda (penalty) such that the MSE is within one standard error of the minimum MSE.

Code

4. Do the same thing for ridge regression.

Select the model with the largest lambda (penalty) such that the MSE is within one standard error of the minimum MSE.

Code

5. HR's decision to have only `NOx` and `RM` enter quadractically seems sort of arbitrary.

Expand the data to contain the square term of all variables. Then run Lasso on this expanded data set. Which coefficients survive now?

Code

6. Report the internal MSE and test data MSE for HR's original model; Lasso and Ridge on the original covariates; and Lasso on the full set of second order terms. Which one fits best?

Code