Tasks

Contents

Introduction

The goal of the midterm exam is to apply some of the methods covered in our course by now to a new dataset. We will work with the data characterizing real estate valuation in New Taipei City, Taiwan that is available at UCI ML repository as well as at this course website on canvas. The overall goal will be to use data modeling approaches to understand which attributes available in the dataset influence real estate valuation the most. The outcome attribute (Y – house price of unit area) is inherently continuous, therefore representing a regression problem.

For more details please see dataset description available at UCI ML or corresponding HTML file on canvas website for this course. For simplicity, clarity and to decrease your dependency on the network reliability and UCI ML or canvas website availability during the time that you will be working on this project you are advised to download data made available on the canvas website to your local folder and work with the local copy. The dataset at UCI ML repository as well as its copy on our course canvas website is provided as an Excel file Real estate valuation data set.xlsx – you can either use `read_excel` *method from R package* `readxl` *to read this Excel file directly or convert it to comma or tab-delimited format in Excel so that you can use* `read.table` *on the resulting file with suitable parameters (and, of course, remember to double check that in the end what you have read into your R environment is what the original Excel file contains).*

Finally, as you will notice, the instructions here are much terser than in the previous problem sets. We expect that you use what you've learned in the class to complete the analysis and draw appropriate conclusions based on the data. The approaches that you are expected to apply here have been exercised in the preceeding weeks – please feel free to consult your submissions and/or official solutions as to how they have been applied to different datasets. As always, if something appears to be unclear, please ask questions – note that we may decide to change your questions to private mode as we see fit, if in our opinion they reveal too many specific details of the problem solution.

Sub-problem 1: load and summarize the data (20 points)

Download and read in the data, produce numerical and graphical summaries of the dataset attributes, decide whether they can be used for modeling in untransformed form or any transformations are justified, comment on correlation structure and whether some of the predictors suggest relationship with the outcome.

Sub-problem 2: multiple linear regression model (25 points)

Using function `lm` *fit model of outcome as linear function of all predictors in the dataset. Present and discuss diagnostic plots. Report 99% confidence intervals for model parameters that are statistically significantly associated with the outcome and discuss directions of those associations. Obtain mean prediction (and corresponding 90% confidence interval) for a new observation with each attribute set to average of the observations in the dataset. Describe evidence for potential collinearity among predictors in the model.*

Sub-problem 3: choose optimal models by exhaustive, forward and backward selection (20 points)

Use `regsubsets` *from library* `leaps` *to choose optimal set of variables for modeling real estate valuation and describe differences and similarities between attributes deemed most important by these approaches.*

Sub-problem 4: optimal model by resampling (20 points)

Use cross-validation or any other resampling strategy of your choice to estimate test error for models with different numbers of variables. Compare and comment on the number of variables deemed optimal by resampling versus those selected by `regsubsets` *in the previous task.*

Sub-problem 5: variable selection by lasso (15 points)

Use regularized approach (i.e. lasso) to model property valuation. Compare resulting models (in terms of number of variables and their effects) to those selected in the previous two tasks (by `regsubsets` *and resampling), comment on differences and similarities among them.*

Extra points problem: using higher order terms (10 points)

Evaluate the impact of adding non-linear terms to the model. Describe which terms, if any, warrant addition to the model and what is the evidence supporting their inclusion. Evaluate, present and discuss the effect of their incorporation on model coefficients and test error estimated by resampling.