STAT 542, Homework 2

September 21, 2017
Due date: Oct 6, 11:59 pm to Compass

Requirements: This homework is a mini-project that focuses on analyzing the bitcoin dataset from Kaggle:

https://www.kaggle.com/sudalairajkumar/cryptocurrencypricehistory

You should submit your report and R code(s), preferably in separate files. Your report should be in PDF/MS Word format. Font size should be 12pt and plots need to be clearly labeled. Your report should include necessary explanations and should not be a simple output file of the R code. The R code should include comments to help our grading process. This homework worth 100 points total. Late submission penalty is 5 points for each day (round up) of delay.

First, download the dataset from Kaggle, and extract the "`bitcoin_dataset.csv`" file. We will use this data only. You are encouraged to read relevant information from the Kaggle website. The goal is to fit **linear models** to predict the outcome variable $Y =$ `btc_market_price` (the second column) from 1/1/2017 to 9/12/2017 (we will **NOT** use the last two rows). The training dataset is **ANY** information prior to 1/1/2017. Note that in this mini-project, you will face missing data, categorical predictors, computational issues, and maybe other problems. Use your best judgement to deal with them. There is no general "best answer". Once you have the data, perform the following:

Question 1: [35 points] In the first analysis, we will ignore the variable `btc_trade_volume` because it contains missing values. Use remaining part in the training dataset to build the model. You should clearly describe how to use the other variables in your report.

a) [15 Points] Fit the best subset selection to the dataset and report the best model of each size.

b) [15 Points] Use C_p, AIC and BIC criteria to select the best model and report the result from each. Apply the fitted models to the testing dataset and report the prediction error $n_{\text{test}}^{-1} \sum_{i \in \text{test}} (\widehat{Y}_i - Y_i)^2$.

c) [15 Points] Redo a) and b) using $\log(1+Y)$ as the outcome. Report the best models. Then for prediction, transform the predicted values into the original scale and report the prediction error of each model.

Question 2: [65 points] This is the biggest project we had so far. We will write our own Lasso regression algorithm in this question, and apply that on our `bitcoin` dataset. Of course, you are not allowed to use any additional R package. We will try to make the code as compact as possible and just enough to get the job done. When you are writing the code, you should have these details in your mind:

- You can actually write your objective function in several different ways. They are all correct, but corresponds to different choice (scale) of the tuning parameter λ, hence slightly different soft thresholding function. So make a choice:

$$f(\boldsymbol{\beta}, \beta_0) = \|\mathbf{y} - \beta_0 - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{1}$$

$$f(\boldsymbol{\beta}, \beta_0) = \frac{1}{n}\|\mathbf{y} - \beta_0 - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2}$$

$$f(\boldsymbol{\beta}, \beta_0) = \frac{1}{2n}\|\mathbf{y} - \beta_0 - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{3}$$

- We will scale the variables into unit variance, and perform the algorithm, so the solution will be different if you don't scale them.
- We will not penalize the intercept term β_0. The easiest way to deal with it is to center both your X and Y and perform the algorithm. After you get all the $\boldsymbol{\beta}$ for each given λ, you can recalculate β_0.
- Don't make your λ too small. The algorithm may not converge.
- Need criteria to stop the iteration if nothing really changes after each iteration.

Part I: [35 points] Complete the Lasso fitting code. To help you navigate through this task, I created my version of the code, and removed certain part of it for you to complete. Please see the HW2.r file, and finish the task. Once you are done, you should include the necessary part to your report and demonstrate that your code is correct. **Note** that if you prefer to write your own code, that is perfectly fine.

Part II: [30 points] Use your finished code to fit the Lasso model to the `bitcoin` dataset. You do not need to perform cross validation to select the best lambda. Simply fit the training data with a sequence of lambda values, then report the testing errors of them on the testing dataset, and report the best model. Use properly labeled graphs if necessary. If you cannot get the code to work properly, use the `glmnet` package to finish this part and indicate this in the report.