

Problem Set 5

Prof. Conlon

Due: 5/5/19

Packages to Install

The packages used this week are

- ggplot2
- xtable (build tables quickly)
- data.table (data tables are computationally efficient and IMHO easier to work with)

Problem 1 (Coding Exercise)

This exercise asks you to implement and assess the performance of the bootstrap for the linear regression model. Suppose you have the linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where,

- $x_i \sim U[0, 2]$
- $\epsilon_i | x_i \sim U[-1, 1]$
- $\beta_0 = \beta_1 = 1$

We ask you to answer the following questions:

- Write a code that generates i.i.d. samples of sizes $n = 10, 50, 200$ from that distribution, computes (1) the least squares estimator for β , (2) the t-ratio for the least squares coefficient β_1 , $t_n = \frac{\hat{\beta}_{1,LS} - 1}{s.e.(\hat{\beta}_{1,LS})}$, and (3) the least square residuals $\hat{\epsilon}_i = y_i - \hat{\beta}_{0,LS} - \hat{\beta}_{1,LS}x_i$
- Write a code for drawing n times at random from the discrete uniform distribution over the estimated residuals $\hat{\epsilon}_1, \dots, \hat{\epsilon}_n$ (i.e. with replacement).
- Use your code from parts (a) and (b) to implement the residual bootstrap - assuming that ϵ_i and x_i are independent - to estimate the 95th percentiles of the respective distributions of $\hat{\beta}_{1,LS}$ and t_n
- Repeat part (a) for sample size $n = 10, 50, 200$ with 200 replications, where you keep the initial draws of x_1, \dots, x_n from part (a) and only generate new residuals from their conditional distribution. Compute $\hat{\beta}_{1,LS}$ and the statistic t_n using 200 independent samples of size n . Use your results to compute a simulated estimate for the 95th percentiles of the respective sampling distributions for $\hat{\beta}_{1,LS}$ and t_n .
- Compare your results from (c) and (d). What do you conclude about the performance of the bootstrap? How does it compare to the 95th percentile of the asymptotic distribution of t_n ?

Problem 2 (Coding Exercise)

This exercise will walk you through a prediction task. I have downloaded data from a peer-to-peer lending platform, Lending Club. The dataset you will work with is: **`lending_club_07_to_11_cleaned.rda`**. Lending Club provides detailed characteristic information regarding loans, both information on the borrower, as well as, the loan itself. Your goal will be to build a model to predict the outcome of a loan, i.e. whether an

individual paid off a loan or did not pay off a loan. In our case, a good outcome is if the loan is fully paid off, a bad outcome is if the loan is charged off.

The target variable for the analysis is **loan_status**, where:

$$loan_status = \begin{cases} 1 & \text{if loan is paid off} \\ 0 & \text{if loan is not paid off} \end{cases}$$

- a. This is going to be a more DIY style exercise, provide a list of the variables you plan to use for the analysis. Give a short discussion for why you excluded other variables.
- b. Regularization is an important step when using an machine learning algorithm, regularize the variables that you have included. Briefly, why is regularization important?
- c. Provide a simple correlational table to give you a sense of the relationship between your covariates. Do you notice any interesting patterns?
- d. Split the dataset into a single test and training set, a simple rule of thumb is an 40/60 split. How did you build these two sets?
- e. Using your training set, run a logistic regression, a random forest, and a gradient boosted random forest. To show your results, present both a measure of misclassification error, accuracy and a confusion matrix.
- f. One easy way to improve model performance is cross-validation. Do a k-fold cross validation, where k=5, using the best performing model from part (e.). Re-report the misclassification error, accuracy and a confusion matrix.