Problem Set 5

Due on May 21

Please note:

- Submit your answers (in Microsoft Word or PDF format) and your code to canvas. Your answer shall be well written. Graph and Table shall be well-formatted. Your code shall be easy for TA to run and check. Your grade will be affected if your code does not provide proper outputs, or it is confusing so that TA cannot run it.

1. In this question, we are going to explore the standard error of OLS estimator. We will compare the standard error based on asymptotic theory (the one reported in lm() result), and the bootstrap standard error, and the true standard error.

a. Let $e \sim N(0,1)$, $x \sim Gamma(2)$, $\beta_0 = 1$, $\beta_1 = 0.5$. Generate a dataset with $n = 200$ observations by

$$y = \beta_0 + \beta_1 x + e.$$

Use set.seed(1) command before the program. This will be the sample used for bootstrap.

b. Compute $\hat{\beta}_1$ by lm(y~x) command and record the standard error. Record it as the standard error based on asymptotic theory.

c. Use bootstrap to do resampling of the dataset. Generate $B = 500$ bootstrap samples. Compute $\hat{\beta}_1$ for each sample. Then compute the standard deviation of these 500 $\hat{\beta}_1$'s. Record it has the bootstrap standard error of $\hat{\beta}_1$.

d. Use the data generating process in part a to generate $S = 500$ samples (each with sample size $n = 200$) and compute $\hat{\beta}_1$ for each simulation. Then compute the standard deviation of these 500 $\hat{\beta}_1$'s. Record it has the "true" standard error of $\hat{\beta}_1$.

e. Let the sample size $n = 100, 200, 500$. Use the methods in part b, c and d to compute thestandard errors of $\hat{\beta}_1$. Fill the following table.

	SE($\hat{\beta}_1$)		
n	Asymptotic	Bootstrap	True
100			
200			
500			

2. Nonparametric regression and classification: Load the data ipod.csv.

a.We want to use number of bidders (BIDRS) to predict PRICE. Use OLS, KNN (with $k = 1, 5, 10, 15$), and nonparametric regression. Illustrate the prediction of these six models in diagrams.

b. Use k-means algorithm to cluster the sample by BIDRS and PRICE. Illustrate the case for $\kappa = 2, 3, 4, 5$.

3. Lasso: Load the data psid.ps5.csv. Let *wagert* be the outcome variable.

a. Run a OLS regression with all explanatory variable. Record R-squared.

b. Run a default lasso regression. Use set.seed(1) command before the program. Plot the coefficient path.

c. Run a lasso with cross-validation. Use the MSE rule to find the best model. Compute the R-squared of the Lasso model. Compute it with the R-squared from OLS.