STAT/Q SCI 403: Introduction to Resampling Method
Spring 2018
Homework 09

Instructions:

- You have to submit all your answers in a single PDF file generated by either LaTeX or *Rmarkdown*.
- You may use the LaTeXtemplate `HW_template.tex` to submit your answer.
- For questions using R, you have to attach your code in the PDF file. If the question ask you to plot something, you need to attach the plot in the PDF as well.
- If the question asks you to show a figure, the clarity of the figure will also be graded.
- The total score of this homework is 8 points.
- Questions with ♠ will be difficult questions.

Questions:

1. In this question, we will focus on the `rock` dataset. In particular, we treat the variable `area` as the response variable Y and the variables `peri` as the covariate X.

   (a) **(1 pt)** Show the scatter plot and attach a regression line from the simple linear regression.

   (b) **(1 pt)** Fit a kernel regression with the Gaussian kernel and smoothing bandwidth $h = 500$. Show the fitted regression curve in the scatter plot.

   (c) **(1 pt)** Now we consider three smoothing bandwidths: $250, 500$, and 1000. In a scatter plot, attach three fitted regression curves.

   (d) **(1 pt)** Use a 3-fold cross validation to show the cross-validation error versus smoothing bandwidth plot. About which value does the smoothing bandwidth attain the minimum cross-validation error?

   (e) **(1 pt)** Using the smoothing bandwidth $h = 500$, apply a bootstrap approach to construct a 95% confidence interval of the regression function.

2. In this question, we will use the same data (`rock`) and the same variable (X =`peri`, Y =`area`) as the previous question. But now we will use a method called *regressogram*

(regression-histogram). Let $(X_1, Y_1), \cdots, (X_n, Y_n)$ be the observed data. We partition the entire range of the covariate X into five regions:

$$B_1 = (0, 1000],\ B_2 = (1000, 2000],\ B_3 = (2000, 3000],\ B_4 = (3000, 4000],\ B_5 = (4000, 5000].$$

For a point x, assume it falls within the bin B_ℓ. Then we first find all the observations whose covariate are within B_ℓ as well. The estimated regression function $\widehat{m}_n(x)$ is the average of the values of response variable of these observations. Namely, let N_ℓ be the number of observations within B_ℓ.

$$\widehat{m}(x) = \frac{\sum_{i:X_i \in B_\ell} Y_i}{N_\ell} = \frac{\sum_{i=1}^n Y_i I(X_i \in B_\ell)}{\sum_{i=1}^n I(X_i \in B_\ell)}$$

(a) **(1 pt)** There is an interesting fact about the regressogram: for any two points x_1, x_2 such that they are both in the same bin (i.e., $x_1, x_2 \in B_\ell$ for some ℓ),

$$\widehat{m}(x_1) = \widehat{m}(x_2).$$

Explain why this is true.

(b) **(1 pt)** In the scatter plot of the data. Attach a regression curve from the regressogram we just construct.

(c) **(1 pt)** If we want to use the bootstrap to construct a confidence interval, will the empirical bootstrap be a good idea? Why or why not?