# Advanced Regression

Rahul Roy and Suryasis Jana

July-December 2024

## 1 Problem Set 5 : Locally Weighted Regression

1. Write the following lines of code in R.

```
set.seed(1234)
X<-runif(100,5,15)
Y<-5*sin(X)+23*cos(X)^2+rnorm(100,0,5)
Sim.1 <- data.frame(X=X,Y=Y)
reg <- function(x){
  5*sin(x)+23*cos(x)^2
}
```

   The function "reg" represents the truth (i.e., the regression function) of the simulated dataset.

2. Divide the X values into parts. Call the first 80 values the training test. Denote them by X.train. The rest of the 20 values constitute the test set. Call them X.test.

3. Divide the Y values in training set (Y.train) and test set (Y.test) in the similar way.

4. Plot the data such that the points corresponding to the training set and the test set are represented with different symbols.

5. Plot the regression function in the same graph.

6. Apply the Locally Weighted Regression algorithm on the training dataset with the following Kernels.

   (a) Gaussian with $\sigma = 0.1, 0.25, 0.5, 1, 2$. (Try at home)

   (b) Uniform Kernel with $h = 0.6, 0.9, 1.2, 1.5$. (Try at home)

   (c) Epanechnikov Kernel with $h = 0.6, 0.9, 1.2, 1, 5$. (Try at home)

   (d) Biweight Kernel with $h = 0.6, 0.9, 1.2, 1.5$.

Do you find any difficulties while using the other kernels except Gaussian? How do you solve this? For each case, plot the estimates in separate graphs. Comment on your findings (i.e., how do the estimates change with changing $h(\sigma)$?)

7. For each case, compute the training and test errors.

8. Repeat the procedure 50 times. Each time, the training set and the test set should be chosen randomly from the given sample. Find the average of the training errors and the testing errors thus obtained for each case and comment.

9. Consider the "Boston" dataset in the "MASS" library in R and let the variable "medv" be the response ($Y$) and the variable "lstat" be the predictor ($X$). Use 22-fold cross-validation method to find the best choice of $\sigma$ for applying the Locally Weighted Regression with Gaussian kernel.

10. (PS 5B -Homework) Perform the analyses for the simulated dataset as well as the Boston dataset using the *loess* function in R.

11. (PS 6 -Homework) Compare the performance of the best performing models from problem sets 2 to 5 with the best performing model of PS 5B for the "Boston" data set and comment.