Statistical Computing: Homework 5

Due on May 26 (Thursday) 8:30am

1. **Competition (must-do: 10%)**

   There are two data files: **train.csv** (sample size 1000) and **test.csv** (sample size 20000). The target variable to be predicted is y and the input variables are $(x_1, x_2) \in [0, 1]^2$. Use the data in train.csv for modeling training. Fill out the test.csv and submit your results for the competition. The submission format should follow the **submission example csv**. Keep your submission filename as test.csv. We will use the **averaged prediction squared errors** as the performance measure to determine your scores. (Smaller is better!)

2. Optimal (extra points: 5%)

   Consider the data set "Hitters" in R (package: ISLR2). Use GP to predict the salary (output variable Y) of baseball players based on the playing records of the players (input variables $\boldsymbol{x}$). In this problem, for simplicity, only take continuous variables as the input variables in the model fittings.

   Your analysis should include the following:

   - randomly split the data into a training data set (80%) and a testing data set (20%)
   - fit a linear model (LM) for the salary based on the training data
   - fit a GP model for the salary based on the training data
   - compute the predictions and their standard errors (se) for the testing data based on your LM fit and GP fit
   - make comments based on your results

   (You may follow the analysis steps demonstrated in the examples of R Lab9-2)