Overview: In probability and statistics, it is important to understand the mean and variance for any random variables. In many applications, it is straightforward to simulate the random variable Y's, but it is often highly non-trivial to characterize the exact distribution of $Y = Y(X_1, X_2)$ including deriving the explicit formulas for the mean and variance of $Y = Y(X_1, X_2)$ explicitly as a function of X_1 and X_2.

Objective: In this , suppose that $Y = Y(X_1, X_2)$ is a random variable whose distribution depends on two independent variables X_1 and X_2, and the objective is to estimate two deterministic functions of X_1 and X_2: one is the mean $\mu(X_1, X_2) = \mathbf{E}(Y)$ and the other is the variance $V(X_1, X_2) = Var(Y)$.

For that purpose, you are provided the observed 200 realizations of the Y's values for some given pairs (X_1, X_2)'s. You are asked to use data mining or machine learning methods that allow us to conveniently predict or approximate the mean and variance of $Y = Y(X_1, X_2)$ as a function of X_1 and X_2. That is, your task is to predict or approximate two values for those given pairs (X_1, X_2) in the testing data set: one for the mean $\mu(X_1, X_2) = \mathbf{E}(Y(X_1, X_2))$ and the other for the variance $V(X_1, X_2) = Var(Y(X_1, X_2))$.

Training data set: In order to help you to develop a reasonable estimation of the mean and variance of $Y = Y(X_1, X_2)$ as deterministic functions of X_1 and X_2, we provide a training data set that is generated as follows. We first choose the uniform design points when $0 \leq X_1 \leq 1$ and $0 \leq X_2 \leq 1$, that is, $x_{1i} = 0.01 * i$ for $i = 0, 1, 2, \ldots, 99$, and $x_{2j} = 0.01 * j$ for $j = 0, 1, 2, \ldots, 99$. Thus there are a total of $100 * 100 = 10^4$ combinations of $(x_{1i},\ x_{2j}\ )$'s, and for each of these 10^4 combinations, we generate 200 independent realizations of the Y variables, denoted by Y_{ijk} for $k = 1, \ldots, 200$.

The corresponding training data, `7406train.csv`, is available from Canvas. Note that this training data set is a $10^4 \times 202$ table. Each row corresponds to one of $100*100 = 10^4$ combinations of (X_1, X_2)'s. The first and second columns are the X_1 and X_2 values, respectively, whereas the remaining 200 columns are the corresponding 200 independent realizations of Y's.

Based on the training data, you are asked to develop an accurate estimation of the functions $\mu(X_1, X_2) = \mathbf{E}(Y)$ and $V(X_1, X_2) = Var(Y)$, as deterministic functions of X_1 and X_2 when $0 \leq X_1 \leq 1$ and $0 \leq X_2 \leq 1$.

To assist you, a limited empirical data analysis (EDA) on the training data is provided in the appendix by using R. Please feel free to modify to other language such as Python, Matlab, etc.

Testing data set: For the purpose of evaluating your proposed estimation models and methods, we choose 50 random design points for X_1 and 50 random design points for X_2. Thus there are a total of $50 * 50 = 2500$ combinations of (X_1, X_2) in the testing data set. You are asked to use your formula to predict $\mu(X_1, X_2) = \mathbf{E}(Y)$ and $V(X_1, X_2) = Var(Y)$ for $Y = Y(X_1, X_2)$ for the $50 * 50 = 2500$ combination of (X_1, X_2) in the testing data (please keep the six digits for your answers).

The exact values of the (X_1, X_2)'s in the testing data set are included in the file `7406test.csv`, which is available from Canvas. You are asked to use your formula to predict $\mu(X_1, X_2) = \mathbf{E}(Y)$ and $V(X_1, X_2) = Var(Y)$ for the $50 * 50 = 2500$ combination of (X_1, X_2) in the testing data (please keep (at least) six digits for your answers).

<u>**Estimation Evaluation Criterion:**</u> In order to evaluate your estimation or prediction, we obtain "true" values $\mu(X_1, X_2) = \mathbf{E}(Y)$ and $V(X_1, X_2) = Var(Y)$ for each combination of (X_1, X_2) in the testing data set, based on the following Monte Carlo simulations (we will not release these true values!).

We first generated 200 random realizations of Y's for each combination of (X_1, X_2) in the testing data set, but we will not release these 200 independent realizations for the testing data. Next, for each given combination of (X_1, X_2), we have 200 realizations of Y's, denoted by $Y_1, \cdots, Y_{200}$, and then we compute the "true" values as

$$\mu^*_{true} = \bar{Y} = \frac{Y_1 + \cdots + Y_{200}}{200} \quad \text{and} \quad V^*_{true} = \hat{Var}(Y) = \frac{1}{200 - 1} \sum_{i=1}^{200} (Y_i - \bar{Y})^2.$$

Your predicted mean or variance functions, say, $\hat{\mu}(X_1, X_2)$ and $\hat{V}(X_1, X_2)$, will then be evaluated as compared to these true values, $\mu^*_{true}(X_1, X_2)$ and $V^*_{true}(X_1, X_2)$:

$$\begin{aligned} MSE_\mu &= \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} (\hat{\mu}(x_{1i}, x_{2j}) - \mu^*_{true}(x_{1i}, x_{2j}))^2 \\ MSE_V &= \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} (\hat{V}(x_{1i}, x_{2j}) - V^*_{true}(x_{1i}, x_{2j}))^2, \end{aligned} \tag{1}$$

where $(I, J) = (50, 50)$ for the testing data.

<u>**Your tasks:**</u> as your solution set to this , you are required to submit two files to Canvas before the deadline:

(a) A .csv file on the required prediction that includes your predicted values for $\mu(X_1, X_2) = \mathbf{E}(Y)$ and $V(X_1, X_2) = Var(Y)$ for the testing data (in 6 digits). Please name your file as "1.YourLastName.YourFirst.Name.csv",

- The submitted csv file in excel must be 2500 × 4 column, and the first two columns must be the exact same as the provided testing data file "`7406test.csv`". The third column should be your estimated mean $\hat{\mu}(X_1, X_2)$, and the fourth column is your estimated variance $\hat{V}(X_1, X_2)$.
- If you want, please round your numerical answers to the six decimal places, e.g., report your estimations as the form of $30.xxxxxx$, but this is optional: in our evaluation process we will use the round function to round your answers to the six decimal before computing MSE.
- Please save your predictions as a 2500*4 data matrix in this .csv file, e.g., **without headers or row/column labels/names**. We will use the computer to auto-read your .csv file and then auto-compute the MSE values in equation (1) for all students, based on the alphabet order of the last/first name, and thus it is important for you to follow this guideline, e.g., without headers or extra columns/rows in the .csv file and name your .csv file as the above form.

(b) A (pdf or docx) file that explains the methods used for the prediction. Please name your file as "2.YourLastName.YourFirstName"

Your written report should be like good journal papers that is concise, clearly explain and justify your proposed models and methods, also see the guidelines on the final report of our course project. Please feel free to use any methods — this is an open-ended problem, and you can either use any standard methods you learned from the class, or develop your estimation by a completely new approach.

Remark:

- If you upload your files multiple times at Canvas, the file names might be renamed automatically by Canvas to "1.YourLastName.YourFirstName.csv-1" or similar. If this occurs, please do not worry, as we will take this into account and correct for you.
- This essentially asks you to build two different models: one is to predict $\hat{\mu}$ and the other is to predict $\hat{V}$. For each model, there are $p = 2$ independent variables (X_1, X_2). Hopefully this high-level viewpoint allows you easily develop models for prediction.
- After your submission, it is useful to double check whether your submitted .csv file has exactly 2500 rows and 4 columns or not, whether it has "NA" or missing values or not. Some typical small mistakes might severely affect your prediction such as having an extra column or more than 2500 rows, or has some "NA" values, since some models will generate a prediction of "NA" if they are unable to produce a prediction.

Grading Policies: The total point is 25 points, which will be graded by the TAs and instructor. There are three components:

- **Prediction accuracy on mean:** 10 points. The smaller MSE_μ in (1) the better. We expect that most students would have their values in the range of $[1.0, 1.5]$. Thus tentatively, "10" if $MSE_\mu \leq 1.20$, "9" if $(1.20, 1.40]$, "8" if $(1.40, 1.60]$, "7" if $(1.60, 1.80]$, "6" if $(1.80, 2.00]$, "5" if $(2.00, 3]$, "4" if $(3, 10]$, "3" if $(10, 20]$, "2" if $(20, 30]$, etc., and we will keep the right to adjust the grading schedule to be more generous if needed.
- **Prediction accuracy on variance:** 10 points. The smaller MSE_V in (1) the better. We expect that most students would have their values in the range of $[500, 600]$. Thus tentatively, "10" if $MSE_V \leq 550$, "9" if $(550, 570]$, "8" if $(570, 590]$, "7" if $(590, 610]$, "6" if $(610, 630]$, "5" if $(630, 650]$, "4" if $(650, 700]$, "3" if $(700, 1000]$, "2" if $(1000, 5000]$, etc., and we will keep the right to adjust the grading schedule to be more generous if needed.
- **Written Report:** 5 points. There are no specific guidelines on this written report, and please feel free to use the commonsense. With that said, we will look at the following aspects. Is the report well-written or easy to read? Is it easy to find the final chosen model or method? Does the report clearly explain how and why to choose the final chosen method? Does the report discuss how to suitably tune parameters in the final chosen model? We plan to assign the grades of this component as follows: "A"- 5, "B"- 4, "C" - 3, "D"-2, "F"- 1, "Not submitted" - 0):

Appendix: Some useful R codes for (A) training dataset, (B) testing dataset, and (C) our auto-grading program.

(A) Empirical Data Analysis of training dataset, which might be useful to inspire you to develop suitable methods for prediction

```
#####
### Read Training Data
## Assume you save the training data in the folder "C:/temp" in your local laptop
traindata <- read.table(file = "C:/temp/7406train.csv", sep=",");
dim(traindata);
## dim=10000*202
## The first two columns are X1 and X2 values, and the last 200 columns are the Y valus

### Some example plots for exploratory data analysis
### please feel free to add more exploratory analysis
X1 <- traindata[,1];
X2 <- traindata[,2];

## compute the empirical estimation of muhat = E(Y) and Vhat = Var(Y)
muhat <- apply(traindata[,3:202], 1, mean);
Vhat  <- apply(traindata[,3:202], 1, var);

## You can construct a dataframe in R that includes all crucial
##    information
data0 = data.frame(X1 = X1, X2=X2, muhat = muhat, Vhat = Vhat);

## we can plot 4 graphs in a single plot
par(mfrow = c(2, 2));
plot(X1, muhat);
plot(X2, muhat);
plot(X1, Vhat);
plot(X2, Vhat);


## Or you can first create an initial plot of one line
##         and then iteratively add the lines
##
##   below is an example to plot X1 vs. muhat for different X2 values
##
## let us reset the plot
dev.off()
##
## now plot the lines one by one for each fixed X2
##
flag <- which(data0$X2 == 0);
plot(data0[flag,1], data0[flag, 3], type="l",
     xlim=range(data0$X1), ylim=range(data0$muhat), xlab="X1", ylab="muhat");
for (j in 1:99){
  flag <- which(data0$X2 == 0.01*j);
  lines(data0[flag,1], data0[flag, 3]);
}

## You can also plot figures for each fixed X1 or for Vhat

### You are essentially asked to build two models based on "data0":
###  one is to predict muhat based on (X1, X2); and
###  the other is to predict  Vhat based on (X1, X2).
```

(B) Read the testing data and write your prediction on the testing data:

```
## Testing Data: first read testing X variables
testX  <- read.table(file = "C:/temp/7406test.csv", sep=",");
dim(testX)
## This should be a 2500*2 matrix

## Next, based on your models, you predict muhat and Vhat for (X1, X2) in textX.


## Suppose that will lead you to have a new data.frame
##   "testdata" with 4 columns, "X1", "X2", "muhat", "Vhat"
## Then you can write them in the csv file as follows:
## (please use your own Last Name and First Name)
write.table(testdata, file="C:/temp/1.LastName.FirstName.csv",
   sep=",",  col.names=F, row.names=F)

## Then you can upload the .csv file to the Canvas
## Note that in your final answers, you essentially add two columns for your estimation of
##     $mu(X1,X2)=E(Y)$ and $V(X1, X2)=Var(Y)$
##  to the testing  X data file "7406test.csv".
## Please save your predictions as a 2500*4 data matrix
##     in a .csv file "without" headers or extra columns/rows.
```

(C) Our auto-grading program on your prediction (this does not affect your prediction, and it is only for those interested students). Also if somehow the auto-grading program failed (e.g., due to inconsistent file names), we will manually compute your prediction, as we want to make sure to have a fair grading to everyone.

```
##### In the auto-grading, we run loops, one loop for each student
#####  In each loop, we first generate the filename as name1 = "1.LastName.FirstName.csv",
#####  Next, we compare your answers with those Monte Carlo based values,
#####     "muhatestMC" and "VhatestMC", which were computed as mentioned

#####
resulttemp <- read.table(file = name1, sep=",");
muhatmp  <- round(resulttemp[,3], 6);  ## Your predicted values for \mu in 6 digits
```

```
Vhatmp   <- round(resulttemp[,4],6);   ## Your predicted value of Vhat in 6 digits
MSEmu   <-  mean((muhatestMC - muhatmp)^2);
MSEV   <-   mean((VhatestMC - Vhatmp)^2);
##### Your technical scores will be based on MSEmu and MSEV values
##### In general, the smaller MSEs, the better.
##### However, there is no universal answer on how small is small.
##### Also it is more difficult to have accurate prediction on Variance than on Mean
##### END #####
```