Homework #3 Due on 11/14/2020

Instructions: While discussion with classmates is allowed and encouraged, please try to work on the homework independently and direct your questions to me. Please interpret your analysis results using concise and clear language and focusing on interesting findings.

Sports Analytics Project

In this project, we consider the 1992 baseball salary data set, which is available from
http://jse.amstat.org/datasets/baseball.dat.txt
This data set (of dimension 337 18) contains salary information (and performance measures) of 337 Major League Baseball players in 1992. Detailed information about this data set can be found at
http://jse.amstat.org/datasets/baseball.txt
The data set contains the following variables:

Table 1: Variable Description for 1992 Baseball Salary Data

Variable	Columns	Description
Y	1 - 4	salary (in thousands of dollars)
X1	6 - 10	Batting average
X2	12 - 16	on-base percentage (OBP)
X3	18 - 20	number of runs
X4	22 - 24	number of hits
X5	26 - 27	number of doubles
X6	29 - 30	number of triples
X7	32 - 33	number of home runs
X8	35 - 37	number of runs batted in (RBI)
X9	39 - 41	number of walks
X10	43 - 45	number of strike-outs
X11	47 - 48	number of stolen bases
X12	50 - 51	number of errors
X13	53	indicator of “free agency eligibility"
X14	55	indicator of “free agent in 1991/2"
X15	57	indicator of “arbitration eligibility"
X16	59	indicator of “arbitration in 1991/2"
ID	61 - 79	player’s name (in quotation marks)

To bring in the data, use the following R commands:

```
baseball <- read.table(file=
    "http://jse.amstat.org/datasets/baseball.dat.txt",
     header = F, col.names=c("salary," "batting. avg", "OBP," "runs," "hits,"
     "doubles," "triples," "homeruns," "RBI," "walks," and "strike. outs",
     "stolen. bases", "errors," "free.agency.elig", "free.agent.91",
     "arb.elig", "arb.91", "name"))
head(baseball)
```

Linear regression will be used to predict a hitter's salary based on his performance variables. Please follow the steps outlined below to process the analysis.

1. Exploratory Data Analysis: First, prepare your data

   (a) Inspect the data and answer these questions: Are there any missing data? Among all the predictors, how many of them are continuous, integer counts, and categorical, respectively?

   (b) Obtain the histograms of salary and the logarithm (natural base) of salary and comment on the two visualizations. Proceed with the log-transformed salary from this step on.

2. Multiple Linear Regression:

   (a) Partition the data randomly into two sets: the training data D_0 and the test data D_1 with a ratio of about 2:1. You can use the code below for the partitioning:

   ```
   set.seed(123)
   # Sample Indexes
   Index = sample(1:nrow(DataBaseball), size = 0.7*nrow(DataBaseball))
   # Splitting Data
   TrainData  = DataBaseball[Index,]
   dim(TrainData)
   TestData = DataBaseball[-Index,]
   dim(TestData)
   ```

   (b) Using the training data D_0, fit a multiple linear regression model called "fit. train".

   (c) Output the necessary fitting results, e.g., selected variables and their corresponding slope parameter estimates.

   (d) Is there a relationship between the response and predictors?

   (e) Using the model (i.e., "fit. train") you developed above, predict using the test data D_1. Output the mean squared error (MSE).