Homework #3 Due on 10/25/2020

Instructions: While discussion with classmates are 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. Remember to include your Python codes and only necessary Python output in an Appendix. Only e-copy (.pdf) submission via Canvas is acceptable.

In this homework, 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 Python commands:

```
baseball = pd.read_table('http://jse.amstat.org/datasets/baseball.dat.txt',
     header = None, sep= "\s+", names =["salary", "batting.avg", "OBP", "runs",
     "hits", "doubles", "triples", "homeruns", "RBI", "walks", "strike.outs",
     "stolen.bases", "errors", "free.agency.elig", "free.agent.91",
     "arb.elig", "arb.91", "name"])
baseball.head()
```

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

1. Exploratory Data Analysis: First prepare your data
   (a) Obtain the histograms of both salary and the logarithm (natural base) of salary and comment. Proceed with the log-transformed salary from this step on.
   (b) 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?
2. Linear Regression:
   (a) Fit a multiple linear regression model using the entire baseball data. Call it `fit_full`. Provide the output from the model and interpret the results – Which of the variables are significant etc.
3. Linear Regression with variable selection:
   (a) Partition the data randomly into two sets: the training data D_0 and the test data D_1 with a ratio about 2:1. Set `random_state = 42`.
   (b) Using the training data D_0, apply two variable selection methods:
      i. Ridge Regression
      ii. The Lasso
   (c) Report the essential steps and/or key quantities involved in each variable selection methods.
   (d) Output the necessary fitting results for each model, e.g., in particular, selected variables and their corresponding slope parameter estimates.
   (e) Apply the models to the test data D_1. Output the mean squared error (MSE). Let's consider the one yielding the minimum MSE as the "best" final model.
4. Refit your "best" final model using the entire data, i.e., $D_0 \cup D_1$, Call it fit_final. Provide the output (i.e. coefficient estimates) from your final model. Compare the results of fit_final with fit_full.