HW #2
BIOSTAT 705 Spring 2024
Due on Feb 22nd

1. Consider data on corn yield Y (bushels/acre) and rainfall X (inches/yr) in six Midwestern states recorded from 1890 to 1927. Data are given in `corn_yield_and_rainfall.csv`.

   a) Fit a multiple regression model with rainfall and rainfall^2 as predictors in the model. Assess model fit with diagnostic Plots of residuals vs. fitted values and normal Q-Q plot of the standardized residuals. Is this model an improvement over the simple regression model? why?

   b) Plot the residuals from the above model vs. year. Is there a pattern in this plot? ie, yield increases with years after adjusting for rainfall.

   c) If there is a pattern, then fit a multiple regression model with rainfall, rainfall^2 and year as predictors. What is the interpretation of estimated coefficient for year? Is this model better than the model in part a? Why?

   d) Is there a multicollinearity issue in part (c)? if yes, which type?

   e) Examine the data in part c for any influential cases and check for multicollinearity.

   f) Refit the model in part c by adding an interaction term rainfall×year. Is the interaction exist? How to interpret this interaction term and the main effects coefficient?

2. Data on Major League Baseball from the 1986 and 1987 seasons are given in (Hitters.csv).

```
Description
Major League Baseball Data from the 1986 and 1987 seasons.
Format
A data frame with 322 observations of major league players on the following 19 variables.
AtBat: Number of times at bat in 1986, Hits: Number of hits in 1986,
HmRun: Number of home runs in 1986, Runs: Number of runs in 1986,
RBI: Number of runs batted in in 1986, Walks: Number of walks in 1986,
Years: Number of years in the major leagues, CAtBat: Number of times at bat during his career,
CHits: Number of hits during his career, CHmRun: Number of home runs during his career,
CRuns: Number of runs during his career, CRBI: Number of runs batted in during his career,
CWalks: Number of walks during his career,
League: A factor with levels A and N indicating player's league at the end of 1986,
Division: A factor with levels E and W indicating player's division at the end of 1986,
PutOuts: Number of put outs in 1986,
Assists: Number of assists in 1986,
Errors: Number of errors in 1986,
Salary (outcome): 1987 annual salary on opening day in thousands of dollars,

Note, there are 59 players with missing salary. Impute missing salary by the mean of non-missing salaries.
Also, due to large difference in salaries, would be more appropriate to model log(Salary) as an outcome.
```

a) Indicate which subset of predictor variables you would recommend as a ’best’ for predicting player salary by examining plots of the following criteria (vs p number of model predictors): 1) adjusted R^2, 2) C_p, 3) AIC_p and 4) BIC_p.

b) Do the four criteria in part (a) identify the same ’best’ subset? What is the total subset models, one would expect?

c) Repeat parts a and b above using data without imputing the salary. How these differ on selecting a ’best’ subset?

d) Instead of imputing missing salary using mean, use multiple-imputation (MI) method in R via ”mice” package with m=25 number of imputations, method=”pmm” (Predictive mean matching) and maxit=20. Repeat parts a and b with MI data.

e) Now, you analyzed the Hitters dataset using i) complete case data, ie no imputation; ii) used mean as a single imputation method; and iii) using MI method, which onw would you recommend? Why?

f) Perform forward, backward and stepwise procedure to identify the 'best' subset of regression model, using an entry to the model level of significance $\alpha = 0.05$ and remained in the model at $\alpha = 0.05$. Summarize your results.
Use MI dataset for this part as well as for the remaining parts below.

g) Perform ridge and LASSO shrinkage procedures to identify the 'best' regression model. For each shrinkage procedure, show plots of estimated coefficients vs. $\log(\lambda)$ (where λ is a tuning parameter) and deviance as well as cross-validation (cv) plots. What is 'best' estimated λ from the cv method? How the estimated regression model coefficients in LASSO differ for those estimated by ridge regression? Provide Which shrinkage procedure would you recommend? why?

h) Perform cv by splitting the dataset into 2/3 as training and 1/3 as validation set, ie 215 vs. 107. How the estimated "best" λ compares the one produced in part g by LASSO? Is the model selected by cv differ from the LASSO in part d? If yes, which one you recommend? Note: for parts g and h, you need to download "glmnet" in R.