ISOM5610 HW1
Due: 23 Nov, 11:59pm

A bank wants to understand how customer banking habits contribute to revenues and profitability. The bank has customer age and bank account information, e.g., whether the customer has a savings account, whether the customer has received bank loans, and other indicators of account activity.

We want to build a model that allows the bank to predict profitability for a given customer. A surrogate for customer profitability available in our data set is the Total Revenue a customer generates through their accounts and transactions. The resulting model will be used to forecast bank revenues and guide the bank in future marketing campaigns.

The data set contains information on 7,420 bank customers (See p.2)

As there are many predictors in the data set, the bank only wants to focus on a small subset of them but not all. So your team need to filter out some not useful predictors.

1. Split the dataset on training (the first 6000 observations) and testing (the reminding observations) sets.
2. Take log transformation on the Total Revenue and take it as the response variable.
3. By using the training data, do the following
   3.1. (Model.1) Fit a regression model on **log(Total Revenue)** by using all predictors and report the resulting model. You will find some estimates to be NA or warning message like 'singular matrix' (depends on which software you are using), try to figure out the reason behind. (Hints: Using correlation). Propose a possible solution and keep using this solution all the way.
   3.2. (Model.2) Refit the model again using the solution of above.
   3.3. (Model.3) Using forward selection with AIC to select predictors. Report the selected predictors and their estimated coefficients.
   3.4. (Model.4) Using PCR to perform reduction by considering:
      Able to explain at least 90% of the variation
      Relatively low MSE (from cross-validation)
   3.5. (Model.5) Apply LASSO to pick a model with 4 predictors. (with 8-folds for the cross-validation)
4. Compute the predictive MSE for all models above (expect Model.1) by using the testing data. Based on selected predictors from the minimum MSE model, rebuild a multiple regression model with the full dataset using least square fit (Model.6). [Note: In case, Model.4 is the minimum MSE model, then use the second minimum MSE to build Model.6.]
5. By Model.6, predict the **Total Revenue** for a typical non-active customer e.g. average in every numerical predictor and 0 for all indicator predictors.
6. Using the result from Model.6, generate an executive report within one page. You are highly recommended to provide some business implications for your manager in the report.

Rev_Total	Total revenue (in $100) generated by the customer over a 6-month period.
Bal_Tota	Total of all account balances (in $100), across all accounts held by the customer.
Offer	An indicator of whether the customer has received a special promotional offer in the previous one-month period. Offer=1 if the offer was received, Offer=0 if it was not.
AGE	The customer's age.
CHQ	Indicator of debit card account activity. CHQ=0 is low (or zero) account activity, CHQ=1 is greater account activity.
CARD	Indicator of credit card account activity. CARD=0 is low or zero account activity, CARD=1 is greater account activity.
SAV1	Indicator of primary savings account activity. SAV1=0 is low or zero account activity, SAV1=1 is greater activity.
LOAN	Indicator of personal loan account activity. LOAN=0 is low or zero account activity, LOAN=1 is greater activity.
MORT	Indicator of mortgage account tier. MORT=0 is lower tier and less important to the bank's portfolio. MORT=1 is higher tier and indicates the account is more important to the bank's portfolio.
INSUR	Indicator of insurance account activity. INSUR=0 is low or zero account activity, INSUR=1 is greater activity.
PENS	Indicator or retirement savings (pension) account tier. PENS=0 is lower balance and less important to bank's portfolio. PENS=1 is higher tier and of more importance to the bank's portfolio.
Check	Indicator of checking account activity. Check=0 is low or zero account activity, Check=1 is greater activity.
CD	Indicator of certificate of deposit account tier**.** CD=0 is lower tier and of less importance to the bank's portfolio. CD=1 is higher tier and of more importance to the bank's portfolio.
MM	Indicator of money market account activity. MM=0 is low or zero account activity, MM=1 is greater activity.
Savings	Indicator of savings accounts (other than primary) activity. Savings=0 is low or zero account activity, Savings=1 is greater activity.
AccountAge	Number of years as a customer of the bank.