

1. Consider the “Data1” data set. Let y be the response variable and V_1, \dots, V_6 be the predictors and note that V_6 is a categorical predictor. (Hint : you can convert V_6 into indicator predictors by the command `model.matrix`.)
 - (a) Find the best linear regression model using the best subset selection method with the **BIC** criteria for the training data. Based on the predictors selected above, include their second-order interaction terms and rebuild the linear regression model using the best subset selection method with **BIC** criteria. Write down the final fitted model, and its training and test mean square errors.
 - (b) Based on the six predictors V_1, \dots, V_6 and their second-order interaction terms, find the *lasso* regression model with the tuning parameter λ which minimizes the 10-fold cross validation error. Write down the fitted *lasso* regression model, the training and test mean square errors.
 - (c) Based on the six predictors V_1, \dots, V_6 and their second-order interaction terms, find the *ridge* regression model with the tuning parameter λ which minimizes the 10-fold cross validation error. Write down the fitted *ridge* regression model, the training and test mean square errors.
 - (d) Based on the results of (a)–(c), which model would you suggest to use for future prediction? Justify your answer.
2. Consider the “Data2” data set. Let cnt be the count-valued response variable and the following predictors:

t1	: temperature	is_holiday	: indicator for holiday
t2	: apparent temperature(體感溫度)	is_weekend	: indicator for weekend
hum	: humidity in percentage	season	: four seasons
wind_speed	: wind speed in km/h	hour	: 0 to 24 hours
weather_code	: category of weather conditions		

- (a) Merge the training and test data to do the following pre-processing:
 - (i) Convert categorical predictors into indicator predictors.
 - (ii) Standardize the predictors so that all the variables are given a mean of zero and a standard deviation of one.
 - (iii) Convert the count-valued response variable cnt into the following binary variable cnt_new whose label is 1 if its value is larger than the median of cnt and is 0 otherwise.
 - (iv) Split the new data sets into training and test sets by the original partition way.
- (b) Use the training set to build the *LDA* classifier function. Find the training and test classification errors.

- (c) Ignore the predictors with standard deviation equal to zero, then using the remaining training set to build the *QDA* classifier function. Find the training and test classification errors.
 - (d) Build the *KNN* classifier based on the training data set (with all predictors). Construct the table for the training error, leave-one-out cross validation error and the test errors for $1 \leq K \leq 10$. Which value of K attains the smallest cross validation classification error? Plot the training, cross validation and test classification errors versus K ($1 \leq K \leq 10$) in a single figure. Make sure to label the training, cross validation and test error curves. (Hint : Use `knn` and `knn.cv`)
 - (e) Find *lasso* logistic regression model for *cnt_new* with the tuning parameter λ minimizing the 10-fold cross validation classification error. What are the training and test errors? (Hint : Use `cv.glmnet()` with `family="binomial"` and predict the class with the maximum probability)
 - (f) Based on the results of (b)–(e), which model would you suggest to use for future prediction? Justify your answer.
3. Consider the “Data3” data set. Let y be the binary response variable and $gene_1, \dots, gene_{5000}, V_1, V_2, V_3, V_4, V_5$ be the predictors.
- (a) First find the 50 most correlated predictors with y based on *gene* variables. Then fit logistic regression model on these 50 predictors for y and let the fitted value of the logistic regression model be a new predictor \hat{y}_{gene} . Refit the logistic model for y on predictors $\hat{y}_{gene}, V_1, V_2, V_3, V_4, V_5$ by the command `glm()`.
Use command `summary` to print out the p-value of the refitted model.
 - (b) First divide the data into 10-folds. For each fold, using training set to find the 50 most correlated predictors with y and fit the logistic regression model on these 50 predictors for y . Then collect the predictions of each validation sets as new predictor \hat{y}_{gene} . Refit the logistic model for y on predictors $\hat{y}_{gene}, V_1, V_2, V_3, V_4, V_5$ by the command `glm()`.
Use `summary` to print out the p-value of the refitted model.
 - (c) Which result of (a) and (b) is more reasonable? Give your reason.