1. We have seen that as the number of features used in a model increase, the training error will necessarily decrease, but the test error may not. We will now explore this in a simulated data set.

(a) Generate a data set with p = 20 features, n = 1,000 observations, and an associated quantitative response vector generated according to the model: Y = Xβ + ϵ,
where β has some elements that are exactly equal to zero. (be sure to use "set.seed") Hint: you may use "rnorm".

(b) Split your data set into a training set containing 900 observations and a test set containing 100 observations.

(c) Perform subset selection (best, forward or backwards) on the training set, and plot the training set MSE associated with the best model of each size.

(d) Plot the test set MSE associated with the best model of each size.

(e) For which model size does the test set MSE take on its minimum value? Comment on your results. If it takes on its minimum value for a model containing only an intercept a model containing all the features, then play around with the way that you are generating the data in (a) until you come up with a scenario in which the test set MSE is minimized for an intermediate model size.

(f) How does the model at which the test set MSE is minimized compare to the true model used to generate the data? Comment on the coefficient values.

(g) Create a plot displaying $\sqrt{\sum_{j=1}^{p}(\beta_j - \hat{\beta}_j^r)^2}$ for a range of values, r, where $\hat{\beta}_j^r$ is the jth coefficient estimate for the best model containing r coefficients. Comment on what you observe. How do these result compare to part D.

2) This question uses the "Weekly" dataset in the ISLR package. The data contains information for weekly returns for 21 years, beginning in 1990 and ending in 2010.
   a) Produce some numerical and graphical summaries of the "Weekly" data. Do there appear to be any patterns?
   b) Use the full data to perform logistic regression with "Direction" as the response and the five lag variables, plus volume, as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? Comment on these.
   c) Compute the "confusion matrix" and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.
   d) Fit the logistic model using a training data period from 1990-2008, with "Lag2" as the only predictor. Compute the confusion matrix, and the overall correct

fraction of predictions for the held out data (that is, the data from 2009 and 2010).

e) Repeat (d) using LDA.
f) Repeat (d) using KNN with k=1.
g) Which method appears to provide the best results?
h) Experiment with different combinations of predictors, including possible transformations and interactions, for each method. Report the variables, method, and associated confusion matrix that appears to provide the best results on the held-out data. Note that you should also experiment with values for K in the kNN classifier.

3) Consider the Diabetes dataset (posted with assignment). Assume the population prior probabilities are estimated using the relative frequencies of the classes in the data.

(a) Produce pairwise scatterplots for all five variables, with different symbols or colors representing the three different classes. Do you see any evidence that the classes may have difference covariance matrices? That they may not be multivariate normal?

(b) Apply linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA). How does the performance of QDA compare to that of LDA in this case?

(c) Suppose an individual has (glucose test/intolerence= 68, insulin test=122, SSPG = 544. Relative weight = 1.86, fasting plasma glucose = 184). To which class does LDA assign this individual? To which class does QDA?