**Q5.**We now examine the differences between LDA and QDA. -If the Bayes decision boundary is linear, do we expect LDA or QDA to perform better on the training set ? On the test set ?

*If the Bayes decision boundary is linear, we expect QDA to perform better on the training set because its higher flexibility may yield a closer fit. On the test set, we expect LDA to perform better than QDA, because QDA could overfit the linearity on the Bayes decision boundary.*

(B)If the Bayes decision boundary is non-linear, do we expect LDA or QDA to perform better on the training set? On the test set ?

*If the Bayes decision boundary is non-linear, we expect QDA to perform better both on the training and test sets.*

(C)In general, as the sample size nn increases, do we expect the test prediction accuracy of QDA relative toLDAtoimprove,decline,orbeunchanged?Why

*Roughly speaking, QDA (which is more flexible than LDA and so has higher variance) is recommended if the training set is very large, so that the variance of the classifier is not a major concern.*

(D) Even if the Bayes decision boundary for a given problem is linear, we will probably achieve a superior test error rate using QDA rather than LDA because QDA is flexible enough to model a linear decision boundary. Justify your answer.-*False. With fewer sample points, the variance from using a more flexible method such as QDA, may lead to overfit, which in turns may lead to an inferior test error rate.*

**Q7.** Suppose that we wish to predict whether a given stock will issue a dividend this year (“Yes” or “No”) based on XX, last year’s percent profit. We examine a large number of companies and discover that the mean value of XX for companies that issued a dividend was X¯¯¯¯=10X¯=10, while the mean for those that didn’t was X¯¯¯¯=0X¯=0. In addition, the variance of XX for these two sets of companies was σ^2=36σ^2=36. Finally, 80% of companies issued dividends. Assuming that XX follows a normal distribution, predict the probability that a company will issue a dividend this year given that its percentage return was X=4X=4 last year.

*It suffices to plug in the parameters and*XX*values in the equation for*pk(x)pk(x)*. We get*

p1(4)=0.8e−(1/72)(4−10)20.8e−(1/72)(4−10)2+0.2e−(1/72)(4−0)2=0.752;p1(4)=0.8e−(1/72)(4−10)20.8e−(1/72)(4−10)2+0.2e−(1/72)(4−0)2=0.752;

*so the probability that a company will issue a dividend this year given that its percentage return was*X=4X=4*last year is*0.7520.752*.*

**Q8**Next we use 1-nearest neighbors (i.e. K=1K=1) and, which method should we prefer to use for classification of new observations ? Why ?

*In the case of KNN with*K=1K=1*, we have a training error rate of*0%0%*because in this case, we have -* P(Y=j|X=xi)=I(yi=j)P(Y=j|X=xi)=I(yi=j)

*which is equal to*11*if*yi=jyi=j*and*00*if not. We do not make any error on the training data within this setting, that explains the*0%0%*training error rate. However, we have an average error rate of*18%18%*wich implies a test error rate of*36%36%*for KNN which is greater than the test error rate for logistic regression of*30%30%*. So, it is better to choose logistic regression because of its lower test error rate.*

**@Q3.** We now review k-fold cross-validation. – a- Explain how k-fold cross-validation is implemented.

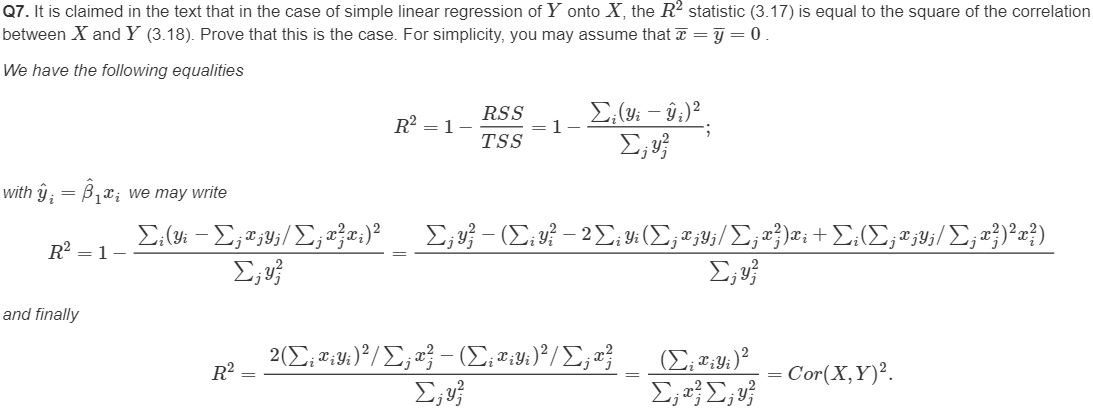
*The k-fold cross validation is implemented by taking the*nn*observations and randomly splitting it into*kk*non-overlapping groups of length of (approximately)*n/kn/k*. These groups acts as a validation set, and the remainder (of length*n−n/kn−n/k*) acts as a training set. The test error is then estimated by averaging the*kk*resulting MSE estimates.* (B)What are the advantages and disadvantages of k-fold cross-validation relative to: (1)The validation set approach ?

*The validation set approach has two main drawbacks compared to k-fold cross-validation. First, the validation estimate of the test error rate can be highly variable (depending on precisely which observations are included in the training set and which observations are included in the validation set). Second, only a subset of the observations are used to fit the model. Since statistical methods tend to perform worse when trained on fewer observations, this suggests that the validation set error rate may tend to overestimate the test error rate for the model fit on the entire data set.* – (2) LOOCV ?

*The LOOCV cross-validation approach is a special case of k-fold cross-validation in which*k=nk=n*. This approach has two drawbacks compared to k-fold cross-validation. First, it requires fitting the potentially computationally expensive model*nn*times compared to k-fold cross-validation which requires the model to be fitted only*kk*times. Second, the LOOCV cross-validation approach may give approximately unbiased estimates of the test error, since each training set contains*n−1n−1*observations; however, this approach has higher variance than k-fold cross-validation (since we are averaging the outputs of*nn*fitted models trained on an almost identical set of observations, these outputs are highly correlated, and the mean of highly correlated quantities has higher variance than less correlated ones). So, there is a bias-variance trade-off associated with the choice of*kk*in k-fold cross-validation; typically using*k=5k=5*or*k=10k=10*yield test error rate estimates that suffer neither from excessively high bias nor from very high variance.*

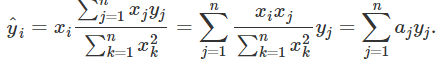
**@Q4.** Suppose that we use some statistical learning method to make a prediction for the response YY for a particular value of the predictor XX. Carefully describe how we might estimate the standard deviation of our prediction.

*We may estimate the standard deviation of our prediction by using the bootstrap method. In this case, rather than obtaining new independant data sets from the population and fitting our model on those data sets, we instead obtain repeated random samples from the original data set. In this case, we perform sampling with replacement*BB*times and then find the corresponding estimates and the standard deviation of those*BB*estimates by using equation*



@Suppose that the true relationship between XX and YY is linear, i.e. Y=β0+β1X+εY=β0+β1X+ε. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.-(a) *knowing more details about the training data, it is difficult to know which training RSS is lower between linear or cubic. However, as the true relationship between*XX*and*YY*is linear, we may expect the least squares line to be close to the true regression line, and consequently the RSS for the linear regression may be lower than for the cubic regression. (B)* Answer (a) using test rather than training RSS. -*In this case the test RSS depends upon the test data, so we have not enough information to conclude. However, we may assume that polynomial regression will have a higher test RSS as the overfit from training would have more error than the linear regression.* (c) Suppose that the true relationship between X and Y is not linear, but we don’t know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.-*Polynomial regression has lower train RSS than the linear fit because of higher flexibility: no matter what the underlying true relationshop is the more flexible model will closer follow points and reduce train RSS. An example of this beahvior is shown on Figure 2.9 from Chapter 2 (D)* Answer (c) using test rather than training RSS - *There is not enough information to tell which test RSS would be lower for either regression given the problem statement is defined as not knowing “how far it is from linear”. If it is closer to linear than cubic, the linear regression test RSS could be lower than the cubic regression test RSS. Or, if it is closer to cubic than linear, the cubic regression test RSS could be lower than the linear regression test RSS. It is dues to bias-variance tradeoff: it is not clear what level of flexibility will fit data better.*

@The KNN classifier is typically used to solve classification problems (those with a qualitative response) by identifying the neighborhood of x0x0 and then estimating the conditional probability P(Y=j|X=x0)P(Y=j|X=x0) for class jj as the fraction of points in the neighborhood whose response values equal jj. The KNN regression method is used to solve regression problems (those with a quantitative response) by again identifying the neighborhood of x0x0 and then estimating f(x0)f(x0) as the average of all the training responses in the neighborhood.  
@If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the best value for K to be large or small? Why? *-* When K becomes larger, we get a smoother boundary, therefore if the boundary is very non linear, we would expect K to be small.

@The sample size nn is extremely large, and the number of predictors pp is small ? -*Better. A flexible method will fit the data closer and with the large sample size, would perform better than an inflexible approach. (B)*The number of predictors pp is extremely large, and the number of observations nn is small ? -*Worse. A flexible method would overfit the small number of observations.* (C)The relationship between the predictors and response is highly non-linear ? - *Better. With more degrees of freedom, a flexible method would fit better than an inflexible one.* (D)The variance of the error terms, i.e. σ2=Var(ε)σ2=Var(ε), is extremely high ? -*Worse. A flexible method would fit to the noise in the error terms and increase variance.*