STAT 636, Fall 2017 - Assignment 4
Due Sunday, Nov. 12, 11:59pm Central

1. Consider the `hof` data. In this problem, you will use both LDA and QDA to build a classifier of HOF status based on the variables `H`, `HR`, `RBI`, `AVG`, `SLG`, and `OBP`. Both LDA and QDA are approximations of the Bayes classifier. Recall that, by default, a Bayes classifier assigns an individual to the class, $\hat{Y}$, for which the posterior probability is maximized:

$$\hat{Y} = \text{argmax}_y \left\{ \widehat{\Pr}\left(Y = y | \text{data}\right) \right\}$$

In this two-class problem, this is equivalent to classifying players as HOF if their posterior probabilities of HOF are greater than 0.5. Alternatively, we could choose to classify players to HOF if their posterior probability of HOF is greater than some arbitrary threshold $\kappa \in [0, 1]$. Consider the following choices for κ:

```
kappa <- seq(from = 0, to = 0.5, by = 0.01)
```

Use leave-one-out cross-validation to assess the performance of the LDA and QDA models for each choice of κ. Specifically, for each value of κ, compute cv-based estimates of the following performance measures:

- sensitivity
- specificity
- positive predictive value (PPV)
- negative predictive value (NPV)
- balanced accuracy, defined in this case as (sensitivity + (3× specificity)) / 4

(a) Report a plot with κ on the horizontal axis and two curves showing cv-based balanced accuracy on the vertical axis; use blue for the LDA curve and green for the QDA curve.

(b) For both LDA and QDA, what are the optimal choices of κ (in terms of balanced accuracy), and what are their corresponding cv-based balanced accuracies, sensitivities, specificities, positive predictive values, and negative predictive values?