CS 5565, ECE 5590CI, CS 465R, HW4(Resampling) 60 pts. (50 pts. UG)

Name ____________________________________

1. (10 points graduate, extra credit UG)
   Using basic statistical properties of the variance, as well as single variable calculus, prove that α given by equation (5.6) in the text book does indeed minimize $\text{Var}(\alpha X + (1-\alpha)Y)$.
   Hints: The following properties will be helpful

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X,Y)$$

$$\text{Var}(aX+b) = a^2\text{Var}(X)$$

$$\text{Cov}(a_1X+b_1, a_2Y+b_2) = a_1a_2\text{Cov}(X,Y)$$

2. (20 points total)
   We will now derive the probability that a given observation is part of a bootstrap sample. Suppose that we obtain a bootstrap sample from a set of n observations.

   (a) (2.5 points) What is the probability that the first bootstrap observation is not the jth observation from the original sample? Justify your answer.

   (b) (2.5 points) What is the probability that the second bootstrap observation is not the jth observation from the original sample?

   (c) (2.5 points) Argue that the probability that the jth observation is not in the bootstrap sample is $(1-1/n)^n$.

   (d) (2.5 points) When n = 5, what is the probability that the jth observation is in the bootstrap sample?

   (e) (2.5 points) When $n = 100$, what is the probability that the jth observation is in the bootstrap sample?

   (f) (2.5 points) When $n = 10,000$, what is the probability that the jth observation is in the bootstrap sample?

   (g) (2.5 points) Create a plot that displays, for each integer value of n from 1 to 100, 000, the probability that the jth observation is in the bootstrap sample. Comment on what you observe.

   (h) (2.5 points) We will now investigate numerically the probability that a bootstrap sample of size $n = 100$ contains the jth observation. Here $j = 4$. We repeatedly create bootstrap samples, and each time we record whether or not the fourth observation is contained in the bootstrap sample.

```
> store = rep(NA, 10000)
> for (i in 1:10000) {
    store[i]=sum(sample(1:100, rep=TRUE)==4) >0
}
> mean(store)
```

   Comment on the results obtained.

3. (20 points total)
   We now review k-fold cross-validation.

   (a) (10 points) Explain how k-fold cross-validation is implemented.

   (b) (10 points) What are the advantages and disadvantages of k-fold cross validation relative to:

      i. The validation set approach?

      ii. LOOCV?

4. (10 points) Suppose that we use some statistical learning method to make a prediction for the response Y for a particular value of the predictor X. Carefully describe how we might estimate the standard deviation of our prediction.