Data Analysis: Assignment 07

17100043 김건하

2.a)

1 - 1/n

2.b)

1 - 1/n

2.c)

부트스트랩에서는 부트스트랩 샘플의 각 관측치가 j번째 관측치과 1/n로 확률이 동일하다.

따라서 n개의 관측치에서 (1-1/n)^n 이 된다

2.d)

1 – (1 – 1/5)^5 = 1 – (4/5)^5 = 0.672

2.e)

1 – (1 – 1/100)^100 = 1 – (99/100)^100 = 0.634

2.f)

1 – (1 – 1/10000)^10000 = 1 – (9999/10000)^10000 = 0.632

3.b)

Validation set approach 는 기존 training data를 두 set로 분할하면 구현할 수 있다.(장점)

하지만 test error rate는 관측치에 따라 매우 달라질 수 있다.(단점)

또한 validation set error rate는 전체 data set에 맞는 모델에 대한 test error rate를 overestimate할 수 있다.(단점)

LOOCV는 k = n 인 경우이다. 따라서 LOOCV는 모델이 n번 적합해야 하므로 계산이 많다.(단점)

또한 LOOCV는 k-fold CV보다 var가 높지만 bias가 낮다.