ISL ASSIGNMENT 4

ANUSHA MUPPALLA

16286311

1) Given, $var(\alpha X+(1-\alpha)Y)$

where given hint, $var(aX+b)=a^2 var(X)$

similarly,

$$var(\alpha X+(1-\alpha)Y) = \alpha^2\sigma_X^2+(1-\alpha)^2\sigma_Y^2+2\alpha(1-\alpha)\sigma_{XY}$$

considering first derivative of

$var(aX+(1-\alpha)Y)$

$$\frac{\partial}{\partial\alpha} var(aX+(1-\alpha)Y) = 2\alpha\sigma_X^2-2\sigma_Y^2+2\alpha\sigma_Y^2+2\sigma_{XY}-4\alpha\sigma_{XY}$$

$$2\alpha\sigma_x^2-2\sigma_y^2+2\alpha\sigma_y^2+2\sigma_{xy}-4\alpha\sigma_{xy}=0$$

$$\alpha = \frac{\sigma_y^2-\sigma_{xy}}{\sigma_x^2+\sigma_y^2-2\sigma_{xy}}$$ (minimum) so

second derivative is

$$\frac{\partial^2}{\partial\alpha^2} var(\alpha x+(1-\alpha)y) = 2\sigma_x^2+2\sigma_y^2-4\sigma_{xy}$$

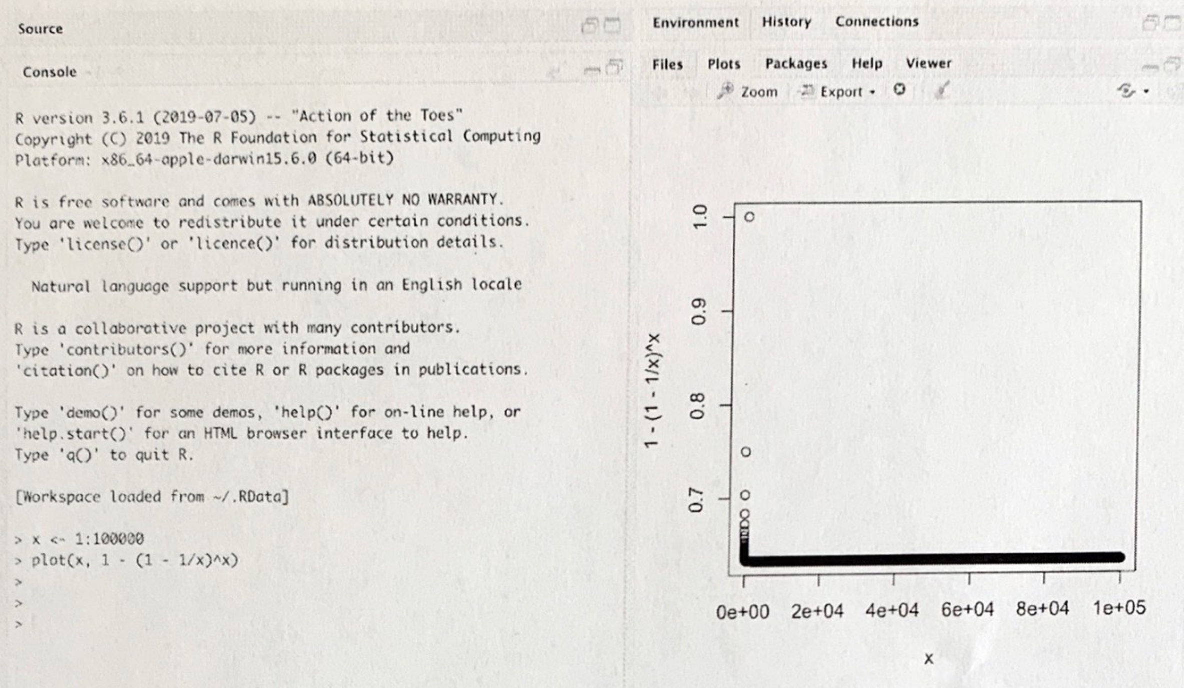
$$= 2\,var(X-Y) \geq 0$$

∴ second derivative is positive

2)

a) The probability that the j^{th} observation is the first bootstrap sample is $1/n$, so the probability is not the first bootstrap sample is $1-1/n$

2)g


Source
Console
R version 3.6.1 (2019-07-05) -- "Action of the Toes"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
[Workspace loaded from ~/.RData]
> x <- 1:100000
> plot(x, 1 - (1 - 1/x)^x)
>
>
>
Environment
History
Connections
Files
Plots
Packages
Help
Viewer
Zoom
Export
1 - (1 - 1/x)^x
1.0
0.9
0.8
0.7
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
x

2)h

Source

Console ~/

```
Platform: x86_64-apple-darwin15.6.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

> x <- 1:100000
> plot(x, 1 - (1 - 1/x)^x)
>
>
> store <- rep(NA, 10000)
> for (i in 1:10000) {
+     store[i] <- sum(sample(1:100, rep = TRUE) == 4) > 0
+ }
> mean(store)
[1] 0.6334
>
```

3)

a) cross validation mean splitting training set into n parts and remaining as test dataset.

the k-fold cross validation is implimented by considering n observations and randomly splitting into k non overlapping groups of length of n/k. this groups as a validation set and the remainder of it acts as training set.

The test error is then estimated by averaging k resulting MSE estimates

b) Advantage of k fold cross validation relative to the validation set: 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 own in the validation set. moreover, error rate may tend to overestimate the test error rate for the model fit on the entire data set

LOOCV :- it is a special case of k-fold cross validation in which k=n

Advantage :-

Advantage of k cross validation relative to LOOC :- it requires fitting the statistical learning method n times. This has the potential to be computationally expensive. moreover, k-fold cv often gives more accurate estimates of the test error rate than does LOOCV

disadvantage :- Bias reduction

4) we may estimate the standard deviation of our prediction by using bootstrap method. in this case, rather than obtaining new independant data sets from the population and fitting our model on those datasets, we instead obtain repeated random samples from the original datasets.

in this case, we perform sampling with replacement B times and then find the corresponding estimates and standard deviation of those B estimates