DAT 500S Module 5

Quiz:

Exercise 5.4: Problem 2 (leave out parts g & h)
Exercise 6.8: Problem 1

Group Assignment:

Exercise 5.4: Problem 2 (part g, h)

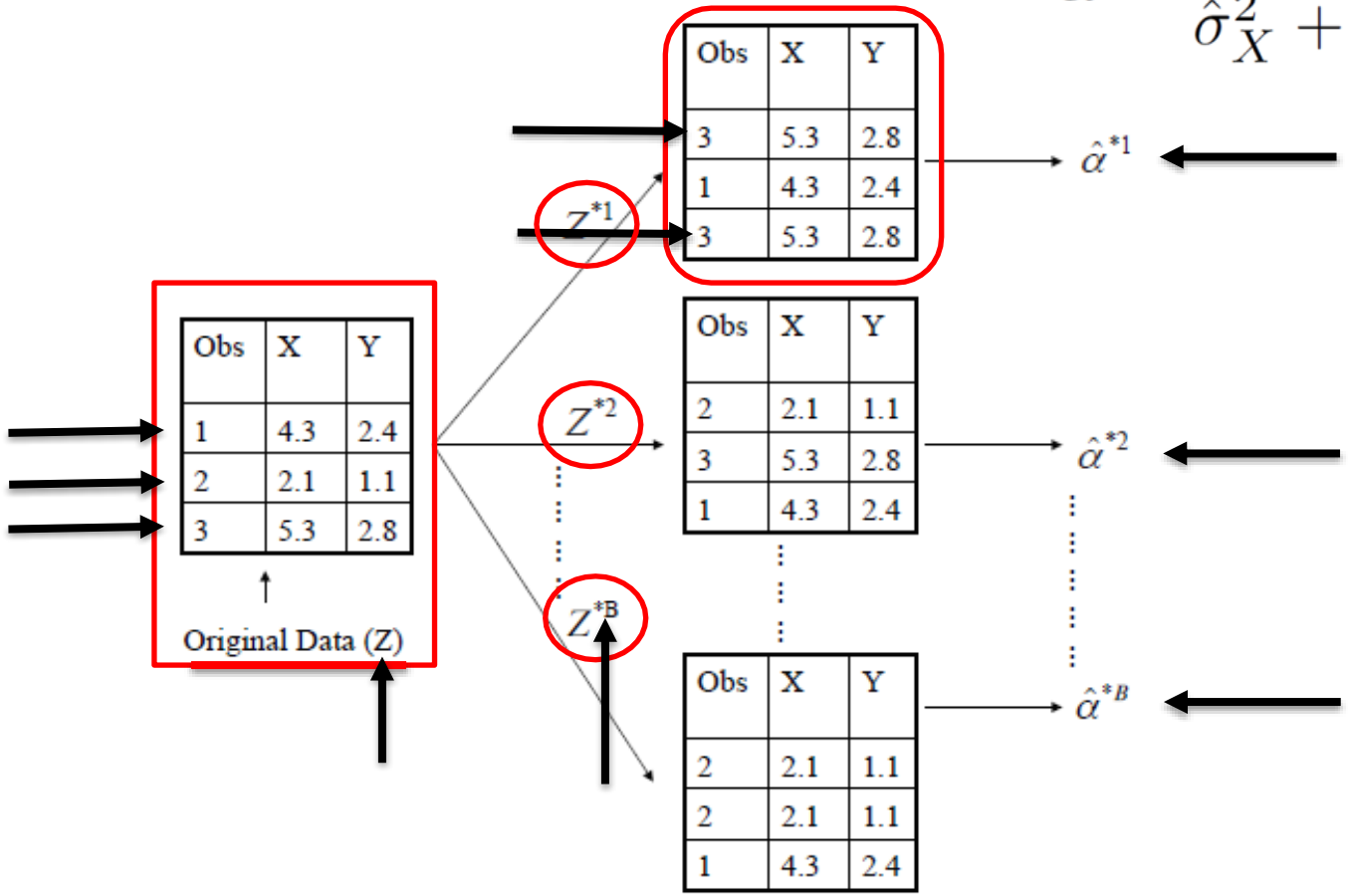
Exercise 5.4: Problem 6

Individual Assignment:

Exercise 6.8: Problem 8 (parts a, b, c, & d)

Module 5 - Bootstrap

$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}$$



Each resampled dataset Z^{*b} is called a *bootstrap replicate*.

Module 5 - Bootstrap

Mean of the alpha's:

$$\bar{\alpha} = \frac{1}{B}\sum_{i=1}^{B} \hat{\alpha}^{*i}$$

Standard Error of the alpha's:

$$SE(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}\left(\hat{\alpha}^{*i} - \bar{\alpha}\right)^2}$$

Exercise 5.4: Problem 2

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) What is the probability that the first bootstrap observation is NOT the j th observation from the original sample? Justify your answer

  1- 1/n,

  Bootstraps are performed with replacement, each observation is picked up independently, so the probability of the first bootstrap observation is the j th observation from the original sample is 1/n.
- (b) What is the probability that the second bootstrap observation is NOT the j th observation from the original sample?

  1- 1/n, same to (a)

Exercise 5.4 Q2

- (c) Argue that the probability that jth observation is NOT in the bootstrap sample is (1-1/n)^n

  each observation in the sample is picked up with replacement

  so, it is independent. ➔ (1-1/n)* (1-1/n)*.... *(1-1/n)

  multiply n times

So, (1-1/n)^n

Exercise 5.4 Q2

- (d) When n = 5, what is the probability that the j th observation is in the bootstrap sample?

  According to c and b:

  n= 5, Probability that j th observation is not in the bootstrap sample: $(1-1/n)^n$ ➔ $(1-1/5)^5 = 0.328$

  So, probability that j th observation is in the bootstrap sample

  1-0.328 = 0.672

- (e) When n= 100, what is the probability that the j th observation is in the bootstrap sample?

  1- (1-1/100)^100 = 0.634

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

  1-(1-1/10000)^10000 = 0.632

Exercise 6.8: Problem 1

We perform best subset, forward stepwise, and backward stepwise selection on a single data set. For each approach, we obtain p + 1 models, containing 0, 1, 2,...,p predictors. Explain your answers:

(a) Which of the three models with k predictors has the smallest training RSS?

Best subset selection will exhibit the smallest training RSS because all possible combinations of predictors are considered for a given k, and it is the best one selected among all k predictors models.

(b) Which of the three models with k predictors has the smallest test RSS?

Best subset selection may have the smallest test RSS, *However, the other methods might also pick a model with smaller test RSS by luck.*

Exercise 6.8: Problem 1

(C) True or False:

i. The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the (k+1)-variable model identified by forward stepwise selection.

True, k+1-variable model identified by forward stepwise selection is performed based on adding one variable to k-variable model identified by forward selection

ii. The predictors in the k-variable model identified by backward stepwise are a subset of the predictors in the (k + 1)- variable model identified by backward stepwise selection.

True, k-variable model identified by backward stepwise selection is performed based on leaving one variable out of the k+1-variable model identified by backward selection

Exercise 6.8: Problem 1

(C) True or False:

iii. The predictors in the k-variable model identified by backward stepwise are a subset of the predictors in the (k + 1)- variable model identified by forward stepwise selection.

False, see (ii)

iv. The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the (k+1)-variable model identified by backward stepwise selection.

False, see(i)

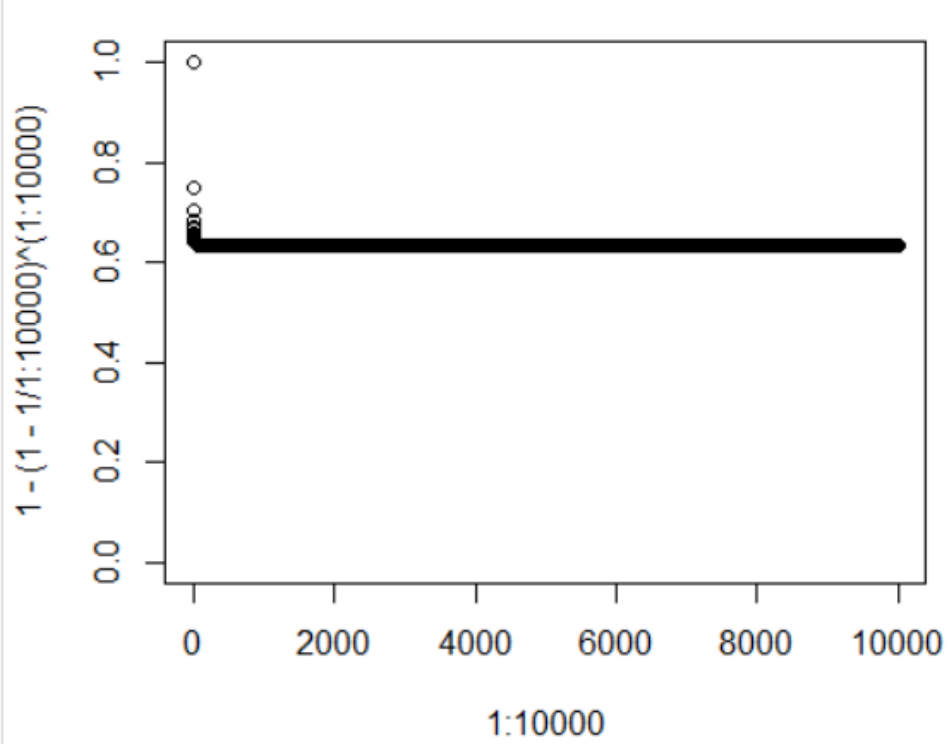
v. The predictors in the k-variable model identified by best subset are a subset of the predictors in the (k + 1)-variable model identified by best subset selection.

False, k-variable model identified by best subset is performed by comparing all combination of k predictors, so k+1 variable model selected by best subset doesn't necessarily contain the combination of k predictors in the k variable model

Group Assignment Exercise 5.4 Q2

(g) 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.

plot(1:10000, 1-(1-1/1:10000)^(1:10000), ylim = range(0,1))



Group Assignment Exercise 5.4 Q2

(h) 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.

Comment:

the resulting fraction of 10,000 bootstrap samples that have the 4th observation is close to our predicted probability of 1-(1-1/100)^100 = 63.4%

Group Assignment Exercise 5.4 Q6

We continue to consider the use of a logistic regression model to predict the probability of default using income and balance on the Default data set. In particular, we will now compute estimates for the standard errors of the income and balance logistic regression coefficients in two different ways:

1. using the bootstrap, and

2. using the standard formula for computing the standard errors in the glm() function. Do not forget to set a random seed before beginning your analysis.

Group Assignment Exercise 5.4 Q6

(a) Using the summary() and glm() functions, determine the estimated standard errors for the coefficients associated with income and balance in a multiple logistic regression model that uses both predictors.

- library(ISLR)
- Default <- na.omit(Default)
- glm.fit = glm(default~income+balance, data= Default,family = binomial)
- summary(glm.fit)

Group Assignment Exercise 5.4 Q6

(b) Write a function, boot.fn(), that takes as input the Default data set as well as an index of the observations, and that outputs the coefficient estimates for income and balance in the multiple logistic regression model.

```
set.seed(1)
boot.fn <- function(df, index) {
  return(coef(glm(default ~ income + balance, data=df, family=binomial, subset=index)))
}
boot.fn(Default, 1:nrow(Default))
```

Group Assignment Exercise 5.4 Q6

(c) Use the boot() function together with your boot.fn() function to estimate the standard errors of the logistic regression coefficients for income and balance.

```
set.seed(1)

library(boot)

boot(Default, boot.fn, 1000)
```

(d) Comment on the estimated standard errors obtained using the glm() function and using your bootstrap function.

Standard error estimates from both methods are similar