ISL Fall 2015 Assignment IV. 100 pts.

NAME(s):

**LOKESHWAR REDDY INJA (16207239/ LIRK9)**

**SANTHOSH KUMAR GATTU (16211118/ SG6N6)**

**SANTOSH KUMAR (16207971/ SK7Z9)**

You may submit this assignment in groups of upto three each. Write your **PIN numbers** on this sheet and include it as the cover page for your submission.

The objective of this assignment is to practice using R, and gain a fundamental understanding of linear regression. Your submission should include both your code as well your answers to the questions.

Electronic submission on Blackboard is due **latest by 11 pm on Wed, Oct 28th**. You may upload upto **three** submissions **before** the deadline – only the last submission will be graded. Submissions received after the deadline will be graded only for effort for a maximum of 70% of the total grade (Refer to class syllabus for detailed grading policy).

**State any assumptions you make, justify your answers, show intermediate steps and explain your results for maximum credit**. All answers should be in your own words with any sources you refer to cited at the appropriate places. Any knowledge you acquire from the Internet should be written in your own words and be appropriately referenced. Copying and pasting from the Internet, each other or any other source will not count as your effort (Refer to class syllabus for detailed policy on plagiarism).

**Remember that answers need to be word-processed (NOT handwritten) and should use R. Submit all your R code as a single merged file for all the assignments.**

Answer the following questions from Chapter 5. Note that the questions are slightly different for graduate and undergraduate students:

Undergraduates: Q2, 5, 8

Graduates: Q2, 6, 8

2.A)

Here we have Bootstrap sample with a set of n observations and since in bootstrap sample we draw the items with replacement. There are n-1 items in the n observations that are not j.

Probability that a given observation is part of bootstrap sample is (1- 1/n)

2.B)

The probability is (1 – 1/n). Here, when we draw an item second time, the set of observations we start with is same since in bootstrap model we draw the items with replacement.

2.C)

The probability that the jth sample is not the first sample in your bootstrap is (1 – 1/n )^n. Total number of observations in bootstrap sample is n. We need to pick n different observations and none of them should be the jth one. Since in bootstrap sample we draw the items with replacement. In this case, we multiply (1 – 1/𝑛 ), n times, hence the answer is (1 – 1/𝑛 )^𝑛 .

2.D)

Given n=5.

The probability that the jth observation is in the bootstrap sample is 1- (1 – 1/𝑛 )

Substituting n=5 in 1- (1 – 1/𝑛 )

=1-[(1-1/5)^5]

= 1-0.8^5

= 0.672

2.E)

Given n=100.

The probability that the jth observation is in the bootstrap sample is 1- (1 – 1/𝑛 )

Substituting n=100 in 1- (1 – 1/𝑛 )

=1-[(1-1/100)^100]

= 1-0.99^100

= 0.624

2.F)

Given n=10,000.

The probability that the jth observation is in the bootstrap sample is 1- (1 – 1/𝑛 )

Substituting n=10,000 in 1- (1 – 1/𝑛 )

=1-[(1-1/10,000)^10,000]

= 1-0.9999^10,000

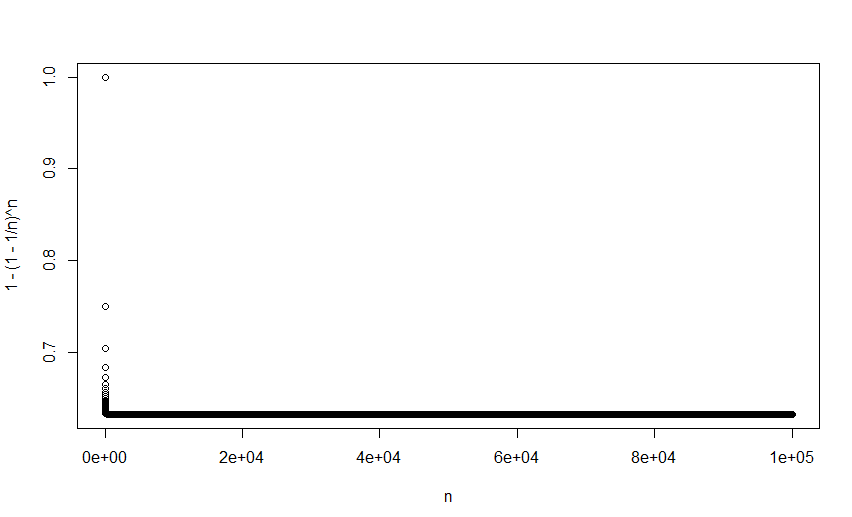
= 1-0.367

=0.633

2.G)

n = 1:100000

plot(n, 1 - (1 - 1/n)^n)



2.H)

> mean(store)

[1] 0.6312

Here we made a list of length 10,000 and each time we sampled observations with size as n=100 with replacement since we are following bootstrap model and we found that out of 100 times 63 times the list contains the number 4.

6.A)

> summary(glm.fit)

Call:

glm(formula = default ~ income + balance, family = "binomial",

data = Default)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.4725 -0.1444 -0.0574 -0.0211 3.7245

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.154e+01 4.348e-01 -26.545 < 2e-16 \*\*\*

income 2.081e-05 4.985e-06 4.174 2.99e-05 \*\*\*

balance 5.647e-03 2.274e-04 24.836 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2920.6 on 9999 degrees of freedom

Residual deviance: 1579.0 on 9997 degrees of freedom

AIC: 1585

Number of Fisher Scoring iterations: 8

The glm() estimates of the standard errors for the coefficients β0, β1 and β2 are respectively 0.4347564, 4.985167210^(-6) and 2.273731410^(-4).

6.B)

boot.fn = function(data, index) {

model = glm(default ~ income + balance, data = data, family = "binomial", subset = index)

return (coef(model))

}

6. C)

ORDINARY NONPARAMETRIC BOOTSTRAP

Call:

boot(data = Default, statistic = boot.fn, R = 1000)

Bootstrap Statistics :

original bias std. error

t1\* -1.154047e+01 -2.176661e-02 4.359958e-01

t2\* 2.080898e-05 -3.506461e-08 4.960480e-06

t3\* 5.647103e-03 1.263640e-05 2.313177e-04

The bootstrap estimates of the standard errors for the coefficients β0, β1 and β2 are respectively 4.359958e-01, 4.960480e-06, 2.313177e-04.

6. D)

The estimates obtained by bootstrap and glm summary are really close to one another.

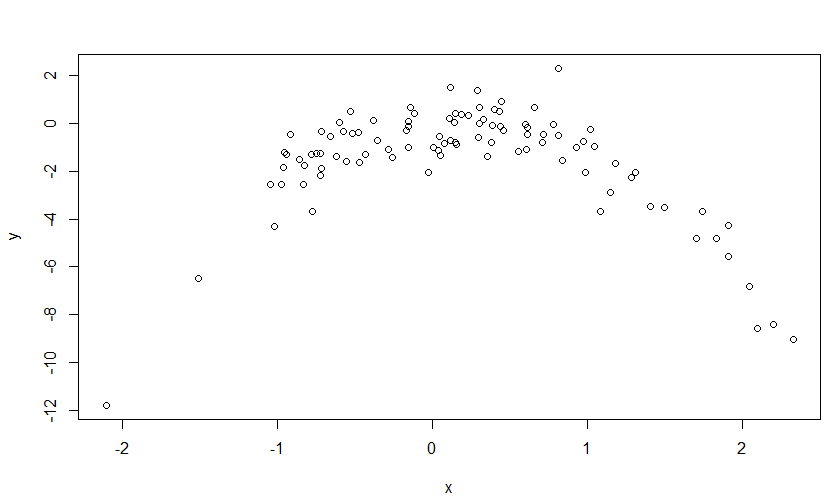
8. A)

n = 100, p = 2

Y=X−2X ^ 2+ ε

8. B)

Plot(x,y)



Quadratic plot ,x from about -2 to 2. y from about -8 to 2.

8.C)

R code file is attached separately.

8. D)

The results are exactly same because LOOCV will be the same since it evaluates n folds of a single observation.

8. E)

The quadratic polynomial had the lowest LOOCV test error rate. This was expected because it matches the true form of y.

8. F)

Call:

glm(formula = y ~ poly(x, 4))

Deviance Residuals:

Min 1Q Median 3Q Max

-2.23392 -0.67400 0.09383 0.46024 3.05683

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.49610 0.08825 -16.952 < 2e-16 \*\*\*

poly(x, 4)1 -5.34771 0.88254 -6.059 2.74e-08 \*\*\*

poly(x, 4)2 -20.52573 0.88254 -23.257 < 2e-16 \*\*\*

poly(x, 4)3 1.04682 0.88254 1.186 0.2385

poly(x, 4)4 -1.84547 0.88254 -2.091 0.0392 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.7788815)

Null deviance: 528.399 on 99 degrees of freedom

Residual deviance: 73.994 on 95 degrees of freedom

AIC: 265.67

Number of Fisher Scoring iterations: 2

p-values show statistical significance of linear and quadratic terms, which agrees with the CV results.