**Homework 2**

**20200639 Chae Woo Jin**

**Chapter 4:** 1, 4, 10, 12, 14

**Chapter 5:** 1, 2, 5, 6, 7

**Chapter 4**

**#1.**

(4.2) , (4.3)

**#4.**

(a) On average, will be used to make the precision, because observations are uniformly distribution.

(b) On average, when , will be used.

(c) On average, when , will be used.

(d) As increases linear, the proportion of observations against whole data set which are used to predict the real value decreases exponentially.

(e) The length of p-dimensional hypercube which contains 10% of the training observation is , so that the volume of this hypercube is , which is a proportion that it contains its training observations.

**#10.**

,

**#12.**

(a) . Therefore, log odds of orange versus apple in my model is

(b) . Therefore, log odds of orange versus apple in friend’s model is

(c) . However, finding a solution of 4 parameters with 2 equations is impossible.

(d) .

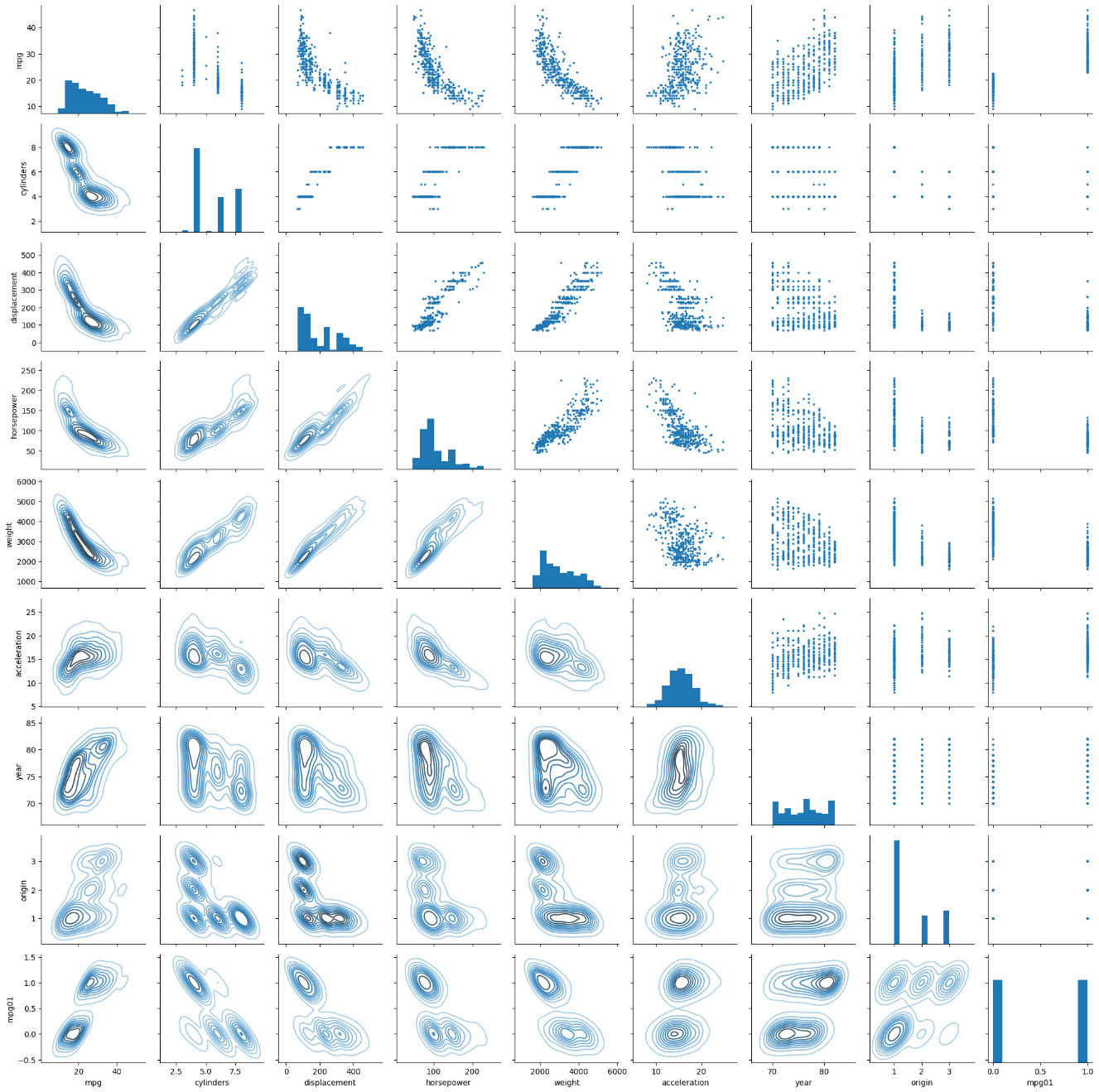
(e) The models are identical with different parameterization so they should perfectly agree.

**#14.**

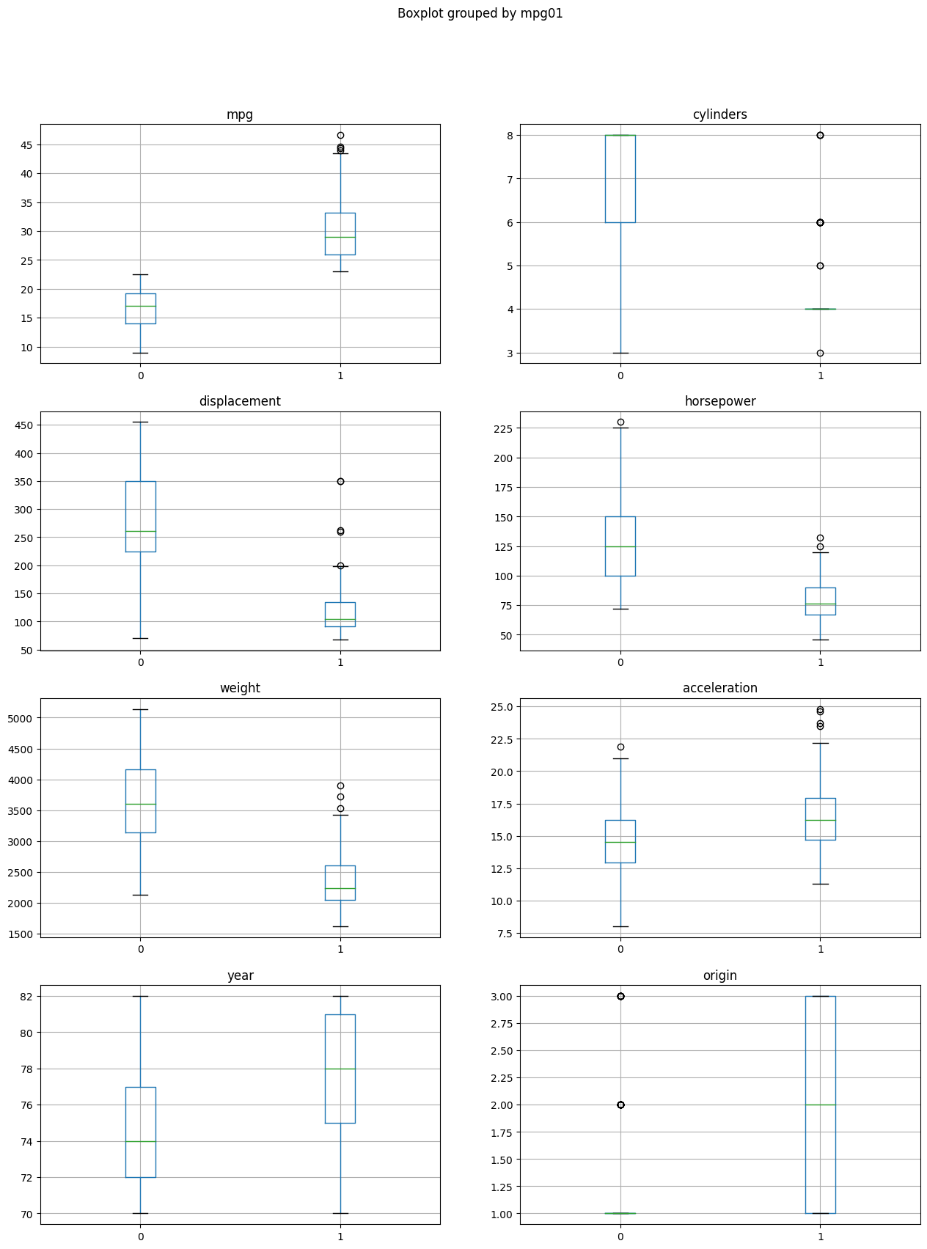
(a)

****

(b)



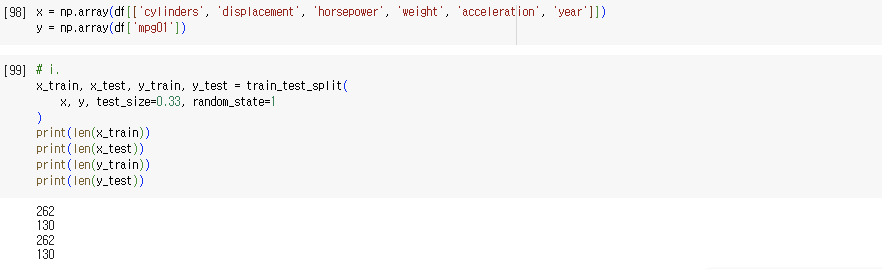
[Scatterplots between features]



[Boxplots grouped by mpg01]

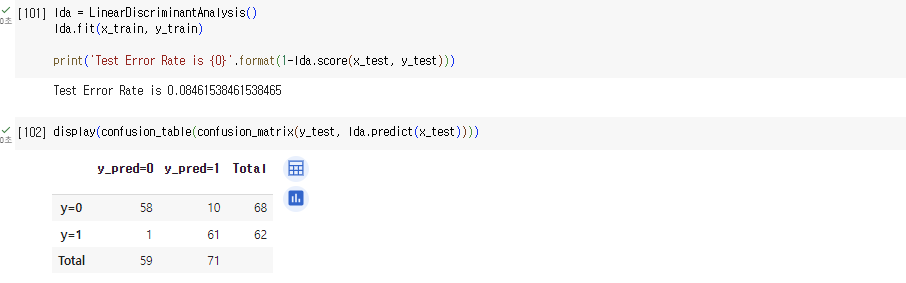
'cylinders', 'displacement', 'horsepower', 'weight', 'acceleration', and 'year' seems well to fit mpg01

(c)



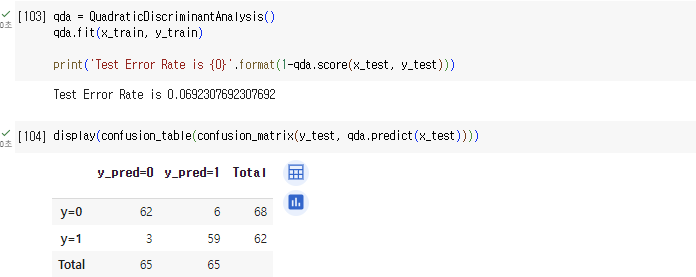
**※ From (d) to (h), 'cylinders', 'displacement', 'horsepower', 'weight', 'acceleration', and 'year' will be used as predictors of ‘mpg01’.**

(d)



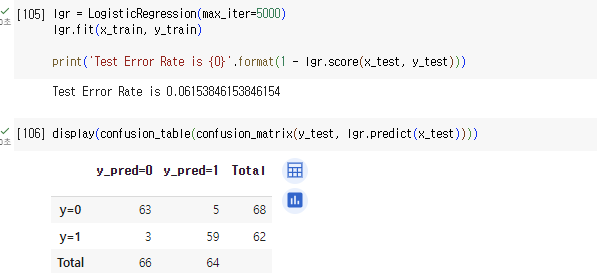
Test error (LDA): 0.084615

(e)



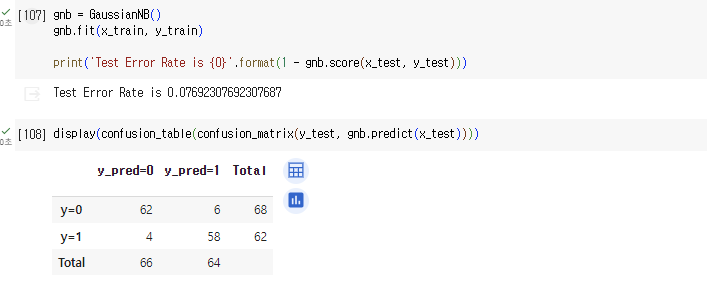
Test error (QDA): 0.069230

(f)



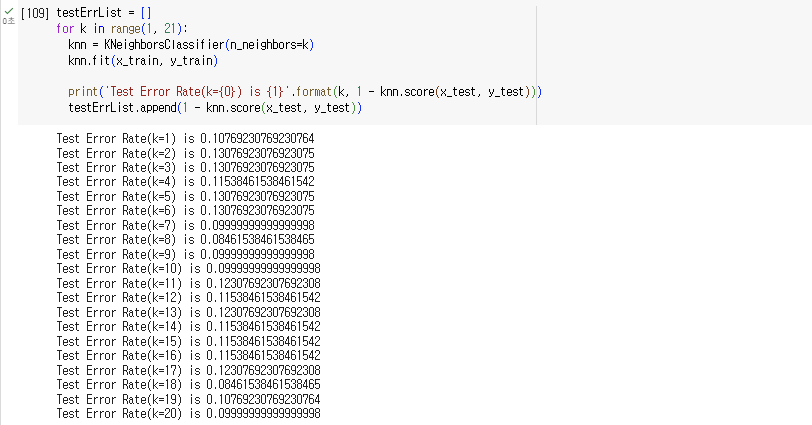
Test error (Logistic Regression): 0.061538

(g)



Test error (Naïve Bayes): 0.076923

(h)



When we consider only overall test error rate, it seems **k=8** perform the best on the data set.

**Chapter 5**

**#1.**

To find that minimizes , lets differentiate it with regard to .

(5.6)

Examining near tells that is minimized at

**#2.**

(a)

(b) As each bootstrap sample is a random sample, this probability is the same.

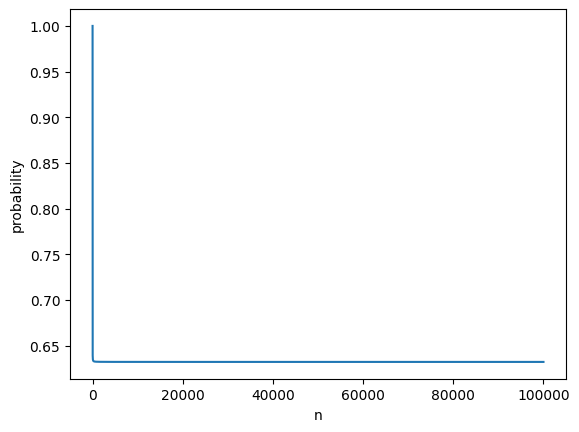
(c) As every sample is independent, it is

(d) It is

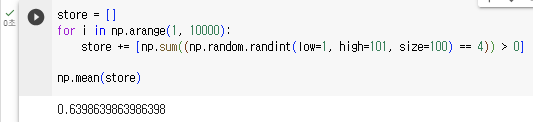
(e) It is

(f) It is

(g) The probability likely to converge in as increases, since .

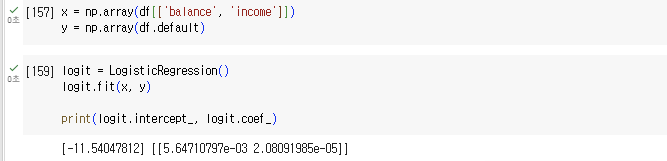


(h) 0.6398639863986398, result from bootstrapping resembles theoretical probability.



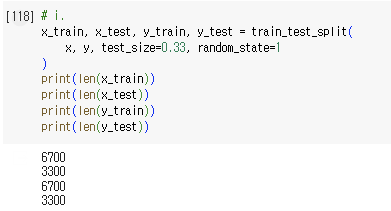
**#5.**

(a)

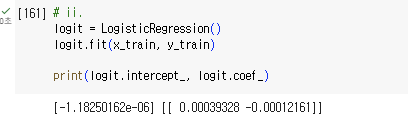


(b)

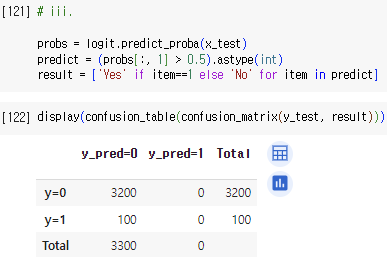
i.



ii.

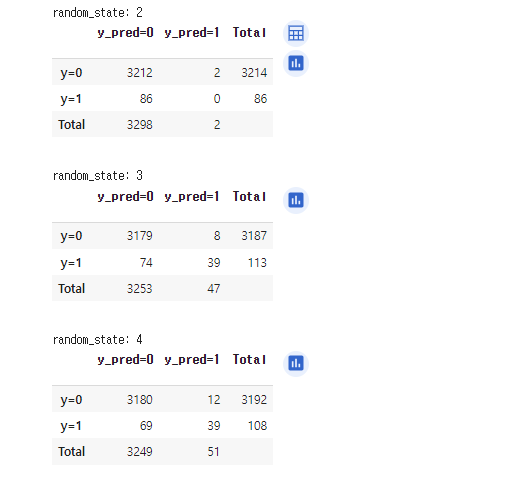


iii.



iv.

Validation set overall error rate is 100/3300 = 0.030303

(c)

**[Validation Set Overall Error Rate]**

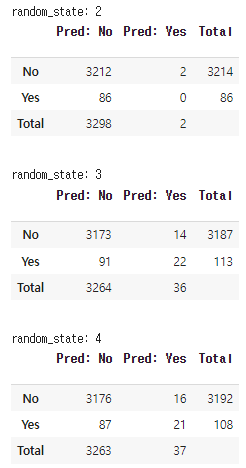
- random\_state (2) => (86+2)/3300 = 0.026667

- random\_state (3) => (74+8)/3300 = 0.024848

- random\_state (4) => (69+12)/3300 = 0.024545

The results obtained are variable and depend on the samples allocated to training vs. test.

(d)

 **[Validation Set Overall Error Rate (Including ‘Student’)]**

- random\_state (2) => (86+2)/3300 = 0.026667

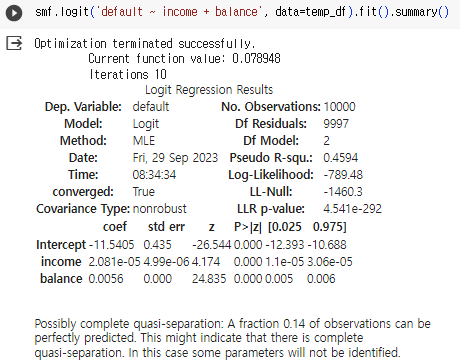
- random\_state (3) => (91+14)/3300 = 0.031818

- random\_state (4) => (87+16)/3300 = 0.031212

Including student does not seem to make a substantial improvement to the test error.

**#6.**

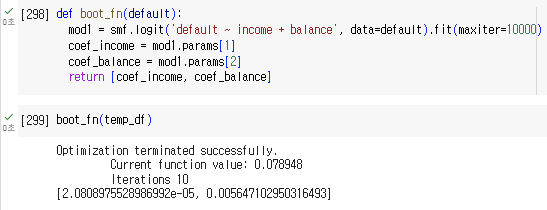
(a)



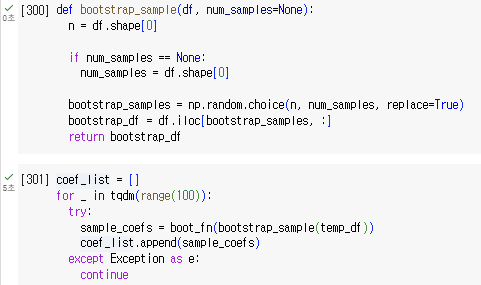
standard error of estimated coefficient(income): 4.99e-06

standard error of estimated coefficient(balance): 0.000

(b)



(c)



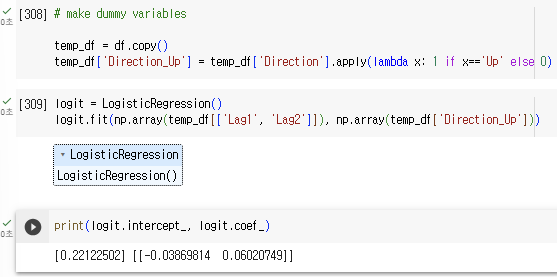
(d)



From (a), estimated coefficients and standard errors of these coefficients for logistic regression of income and balance was [2.081e-05, 0.0056], and [4.99e-06, 0.000] respectively. According to the bootstrapping method, we obtained coefficients [0.000020, 0.005639], standard errors [0.000006, 0.000228] for these coefficients. From this result, one can conclude that result from bootstrapping resembles well with the original one.

**#7.**

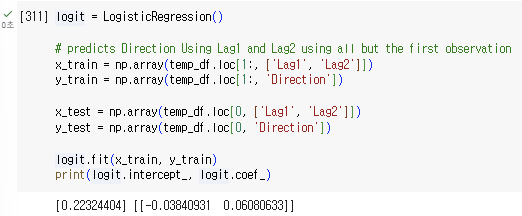
(a)



estimated coefficient (Lag1): -0.03869814

estimated coefficient (Lag2): 0.06020749

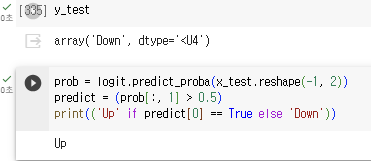
(b)



estimated coefficient (Lag1): -0.03840931

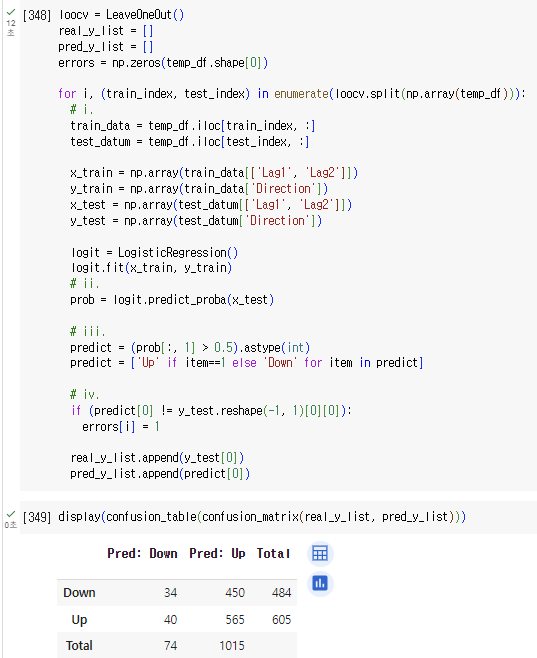
estimated coefficient (Lag2): 0.06080633

(c)

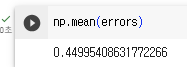


It isn't correctly classified.

(d)



(e)



The LOOCV test error rate is 45% which implies that our predictions are marginally more often correct than not.