**Data 558: Statistical Machine Learning**

**Spring 2023**

**Homework – 3**

**Arjun Sharma**

**Conceptual Questions:**

1. Problem 1:

Review k-fold cross validation. Specifically:

(a) explain how k-fold cross-validation is implemented,

(b) advantages and disadvantages of k-fold cross-validation relative to validation set approach and LOOCV.

Answer:

(a) In K-fold cross validation, the data available to the user is randomly split into ‘K’ smaller, equally sized subsets, also referred to as *folds*. The first fold is treated as the validation set, and the model is fit on the remaining K-1 folds of data. The process is repeated K times where K ϵ *N*. Each time, the data is randomly sampled. The metric used to evaluate model performance is usually an error metric, typically the Mean-Squared Error, which is calculated for the model performance on the test sets. At the end of the process, we obtain K values of the test-error. The K-fold cross validation estimate is computed by averaging the said values. Mathematically, this is expressed as:

CV(K) = MSE1 + MSE2 + MSE3 + … + MSEK =

(b)

**K-Fold Cross Validation vs. LOOCV**

Advantages of K-Fold CV vs LOOCV:

1. Much lower number of repetitions of fitting and evaluating the model. Computationally very inexpensive compared to LOOCV, especially for more complex models.

2. The MSE for K-fold cross validation is much less prone to variance, a model fit using K-fold cross validation can be expected to be more generalizable than one trained using LOOCV

3. K-Fold CV gives better estimates of test error rates than LOOCV. This can be attributed to the bias-variance trade-off, as models that are fitted using the K-Fold CV techniques are less prone to variance compared to models fitted using LOOCV, which uses one datapoint as a validation set for each iteration.

Disadvantages of K-Fold CV vs LOOCV:

1. There is no randomness involved in fitting a model using the LOOCV technique, in contrast, the random selection of training examples in each fold in k-fold cross validation leaves room for some variability in the testing estimate, which is not the case in LOOCV.

2. In situations where we are interested in locating the minimum point in the estimated test MSE curve, K-Fold cross validation cannot be precisely calculated, however, this is possible when LOOCV is employed.

3. LOOCV is much less prone to bias because the validation set in each iteration is a singular datapoint.

**K-Fold Cross Validation vs Validation Set Approach**

Advantages of K-Fold CV vs Validation Set Approach:

1. In the case of the validation set approach 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 in the validation set. The K-fold cross validation techniques – by iterating through the dataset K times – helps reduce this variance in test error rate.

2. In the validation approach, only a subset of the observations—those that are included in the training set rather than in the validation set—are used to fit the model. The k-fold cross validation approach allows for the model to work with a larger number of datapoints because of the randomness of the sampling and repetition of the same.

Disadvantages of K-Fold CV vs Validation Set Approach:

1. K-Fold Cross Validation is potentially computationally more expensive than validation set approach because of repetition of the training process, this is exacerbated in case of more complex models.

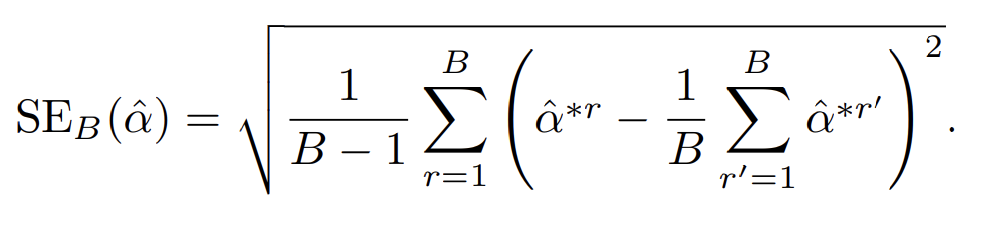
2. Since we average our resulting test-error estimates from the k testing sets in the K-Fold CV approach, we may obtain a relatively ambiguous test-error estimate using this approach compared to the validation set approach, which gives us the test-error estimate resulting from the singular test set.

2. Problem 2:

When do you expect the bootstrap procedure to give accurate confidence intervals and when would it give very inaccurate confidence intervals?

Answer:

The error in the bootstrap estimates can be computed mathematically based on the equation below:



Where B is the number of bootstraps, and α^\*r is the estimate for α.

The judgement for when bootstrapping procedure gives accurate confidence intervals comes down to minimizing the standard error.

Bootstrapping gives us accurate confidence intervals when:

1. B is large. From the Standard Error equation above, Larger B values would be less sensitive to influence of bootstrapped datasets which have extreme datapoints, and

2. Size of each bootstrapped dataset (i.e. n for each Z) is also large. Smaller bootstrapped datasets would be prone to influence from extreme values.

Bootstrapping gives us inaccurate confidence intervals when:

1. B is small. From the Standard Error equation above, smaller values of B would lead to risk of higher impact due to bootstrapped datasets having extreme datapoints, and

2. Size of each bootstrapped dataset is small. In this scenario, extreme values in a smaller dataset can lead to more noise.

3. Problem 3:

What is cross-validation trying to mimic? Describe two scenarios where cross-validation may not be appropriate. Please give a complete description

Answer:

K-fold cross validation attempts to mimic the process of dividing the available data into training and testing data in order to evaluate how the model has generalized on the data, and how the model is expected to perform on real-world data upon deployment.

In other words, the K-fold cross validation process is an imitation of the validation set approach. However, the approach is not always appropriate. Here are some scenarios where K-fold cross validation fails, or is at best a sub-optimal approach:

1. Time-Series Problems:

In time series problems, when the data is distributed over k-folds at random, the chronology of the datapoints would get disturbed. The model might end up being tested on some datapoints that occurred earlier than some of the datapoints it was trained on. For example, consider a model being developed to predict heart attacks based on the user’s blood pressure and pulse data over time, it is possible that the use of k-fold cross validation would lead to the model being trained on data from January 1, 2023, while being tested on data from say, November 25, 2022.

This is a problem because:

a. The order of time being jumbled leads to erratic trends over time and,

b. Data leakage occurs when the latest data and older data are used together.

2. Problems with imbalanced data:

When k-fold cross validation is employed on imbalanced data either by the outcome or by the data distribution, it may introduce bias, since the data may not be properly distributed across the different folds of data. Going back to the example above, we would expect the instances of a datapoint corresponding to a heart attack to be very rare compared to normal readings. If K-fold cross validation were to be applied to such data, it is possible to have imbalanced folds, some of the situations below could occur:

a. Some folds having no datapoints corresponding to a heart attack,

b. Some folds having much more datapoints corresponding to a heart attack compared to other folds, etc.,

In scenarios like these, K-fold cross validation is not a suitable approach.

**Applied Questions:**

Problem 1:

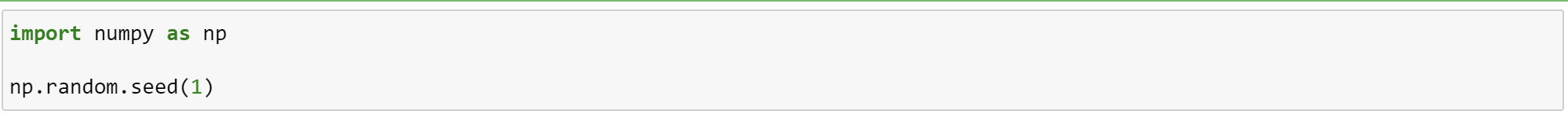
In this problem, you will code your own cross-validation procedure and test it out on simulated data. Generate the simulated data as follows. Let x be a standard normal random variable and y = 0.5 + 0.5x − x2 + x3 + ε where ε is also a standard normal. Draw n = 100 samples that will be used for training and validation, and another n = 100 samples that will be used for evaluating test error.

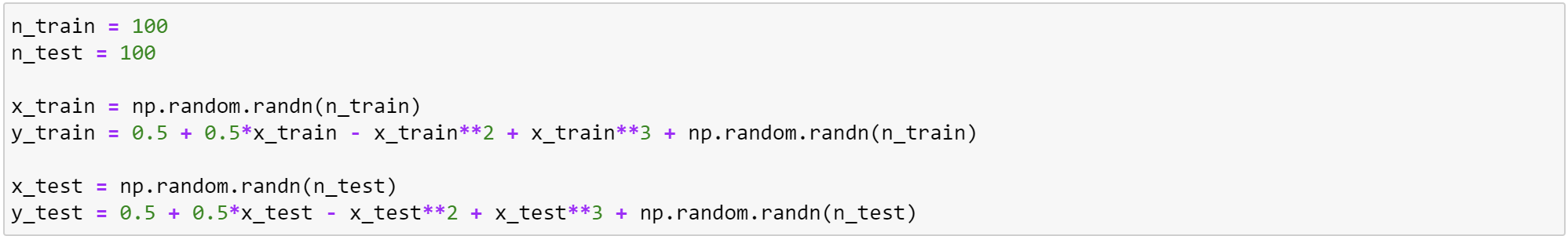
(a) (10 points) Perform holdout validation on linear, quadratic, cubic, quartic, and quintic func-  
tions. Produce a plot with x-axis being complexity, y-axis MSE, and curve of validated  
MSE and the true test MSE. [Here, the true test MSE is the MSE obtained by fitting the a  
given complexity model to the test data]. Repeat the same steps for 10-fold validation and  
leave-one-out validation.

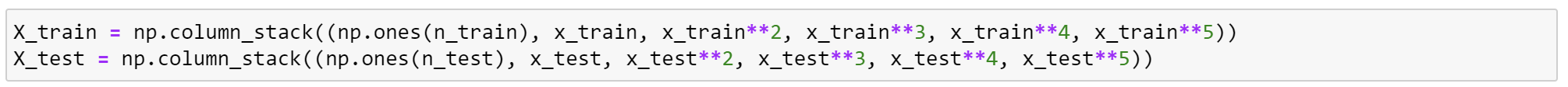
(b) (6 points) What is the optimal complexity chosen by each validation approach? What about  
the corresponding validation MSE? Which validation approach performs most favorably?

(c) (6 points) Repeat step (a) but with seed numbers 5 and 10. What do you notice?

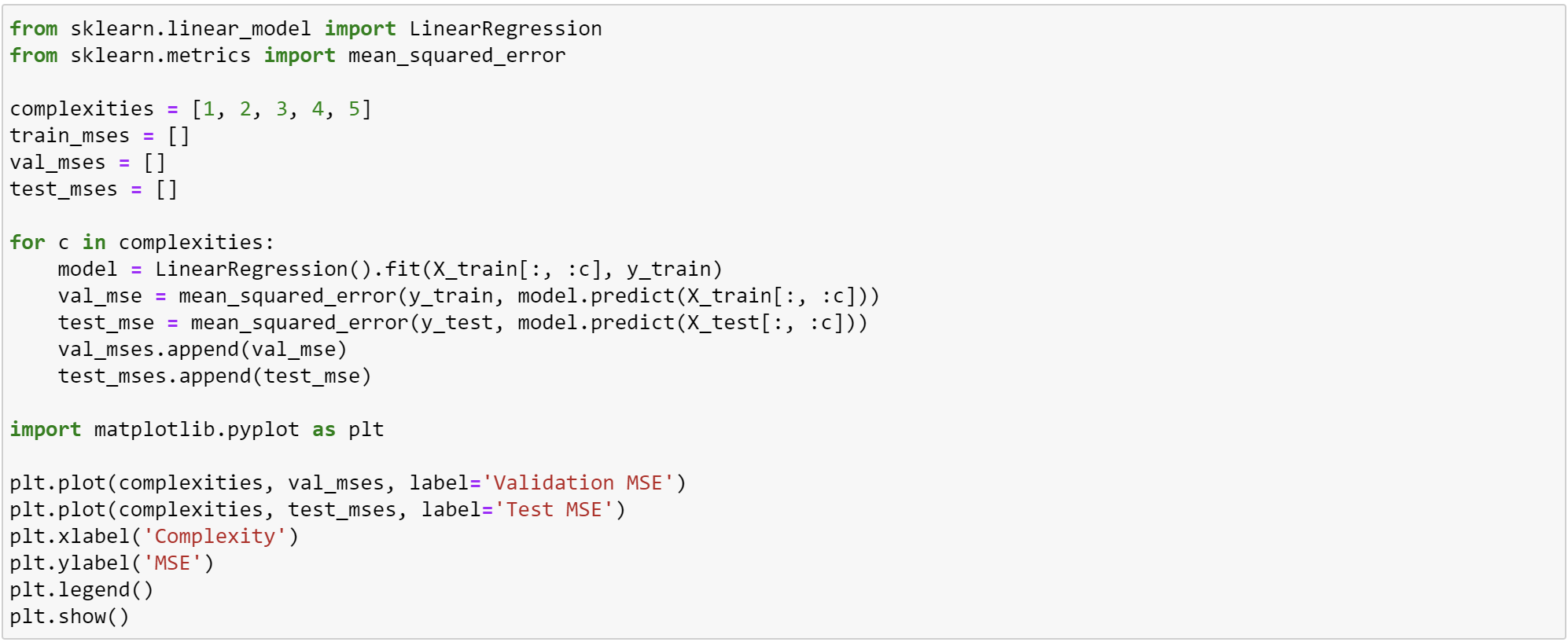
(a)

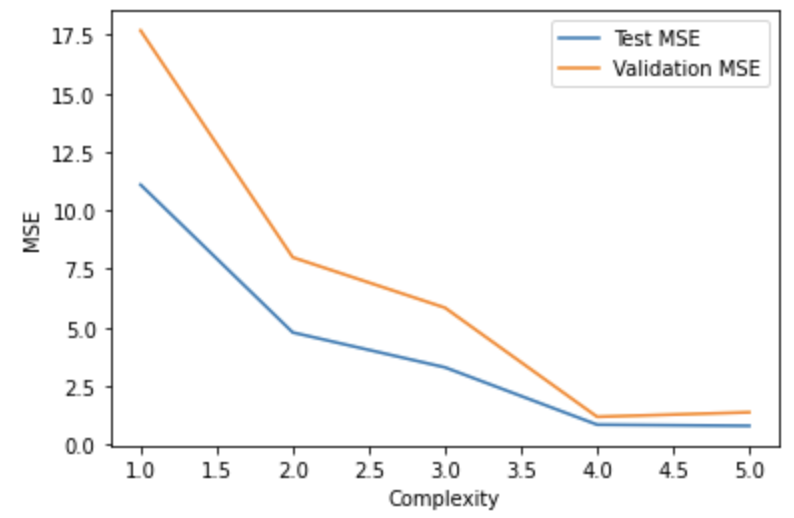


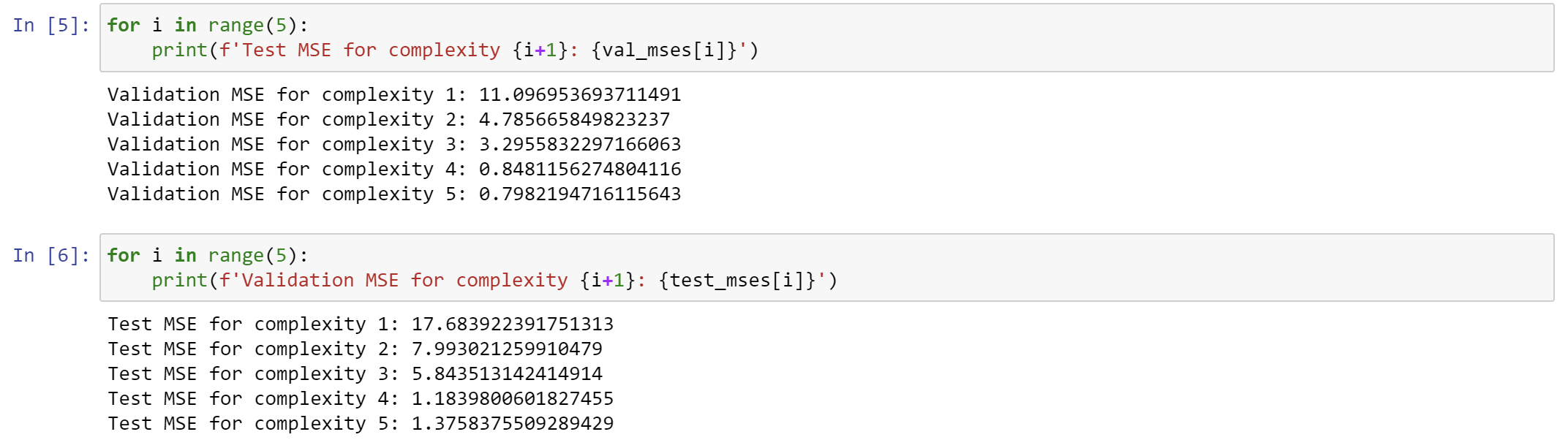




Holdout Validation

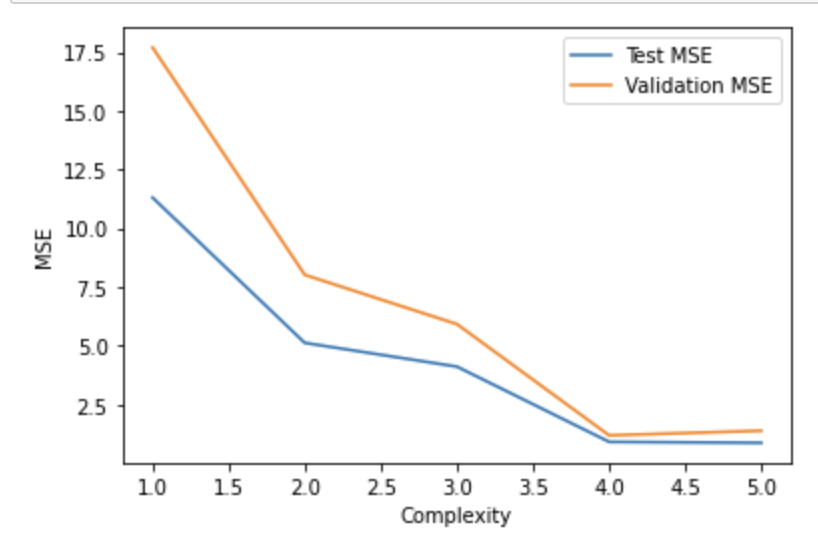


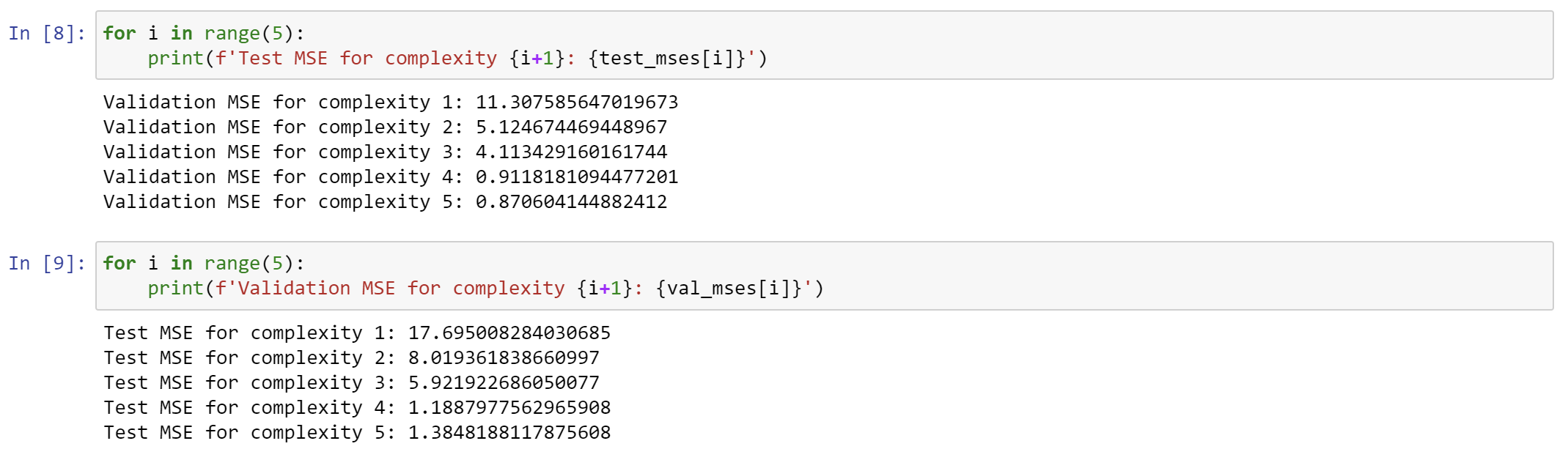




K-Fold Cross Validation



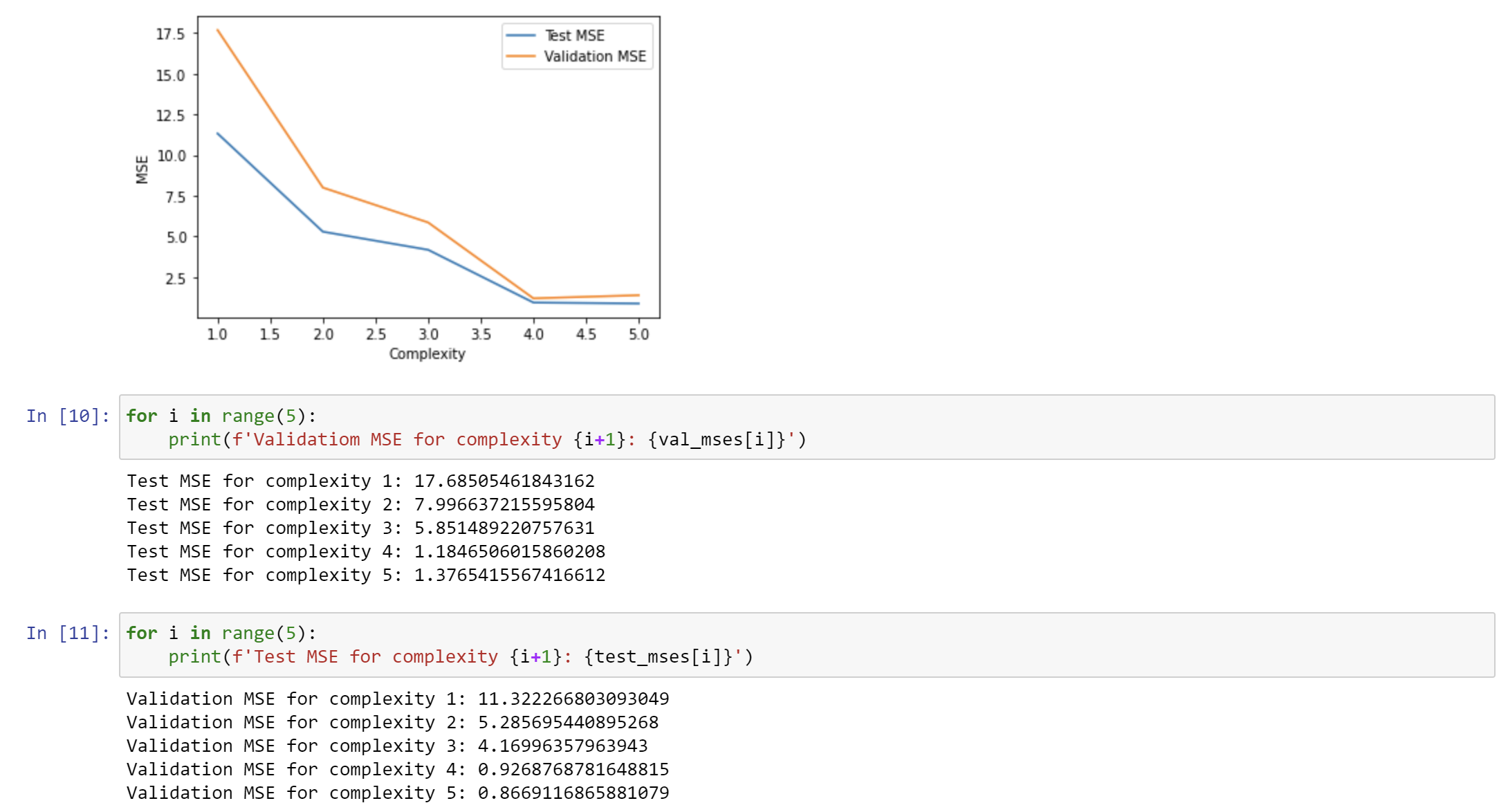




LOOCV

The code for K-fold cross validation is used, except that K = number of data points.

Results:



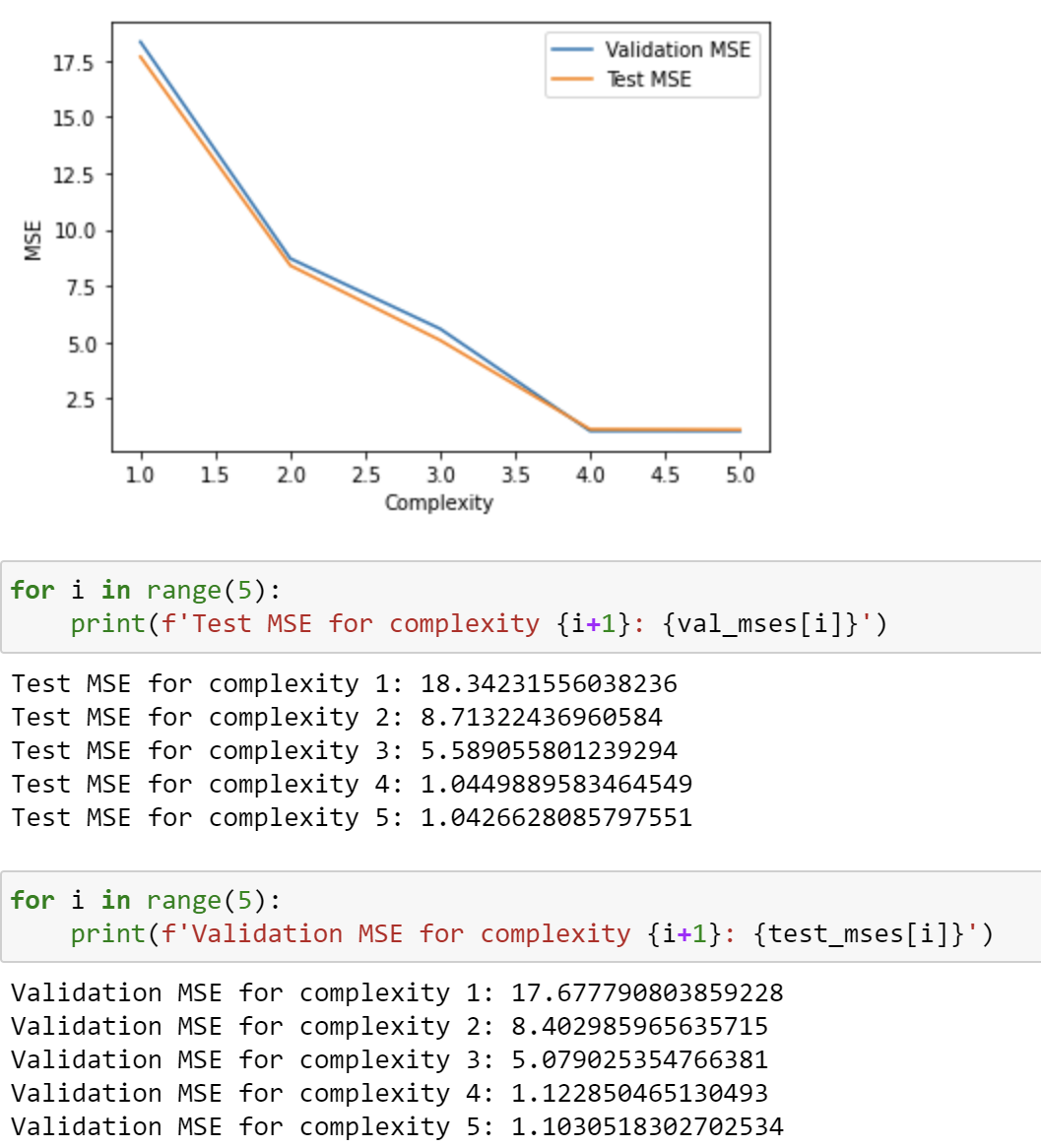
(b)

The optimal complexity chosen by all approaches is 5. LOOCV approach performs the best out of all three.

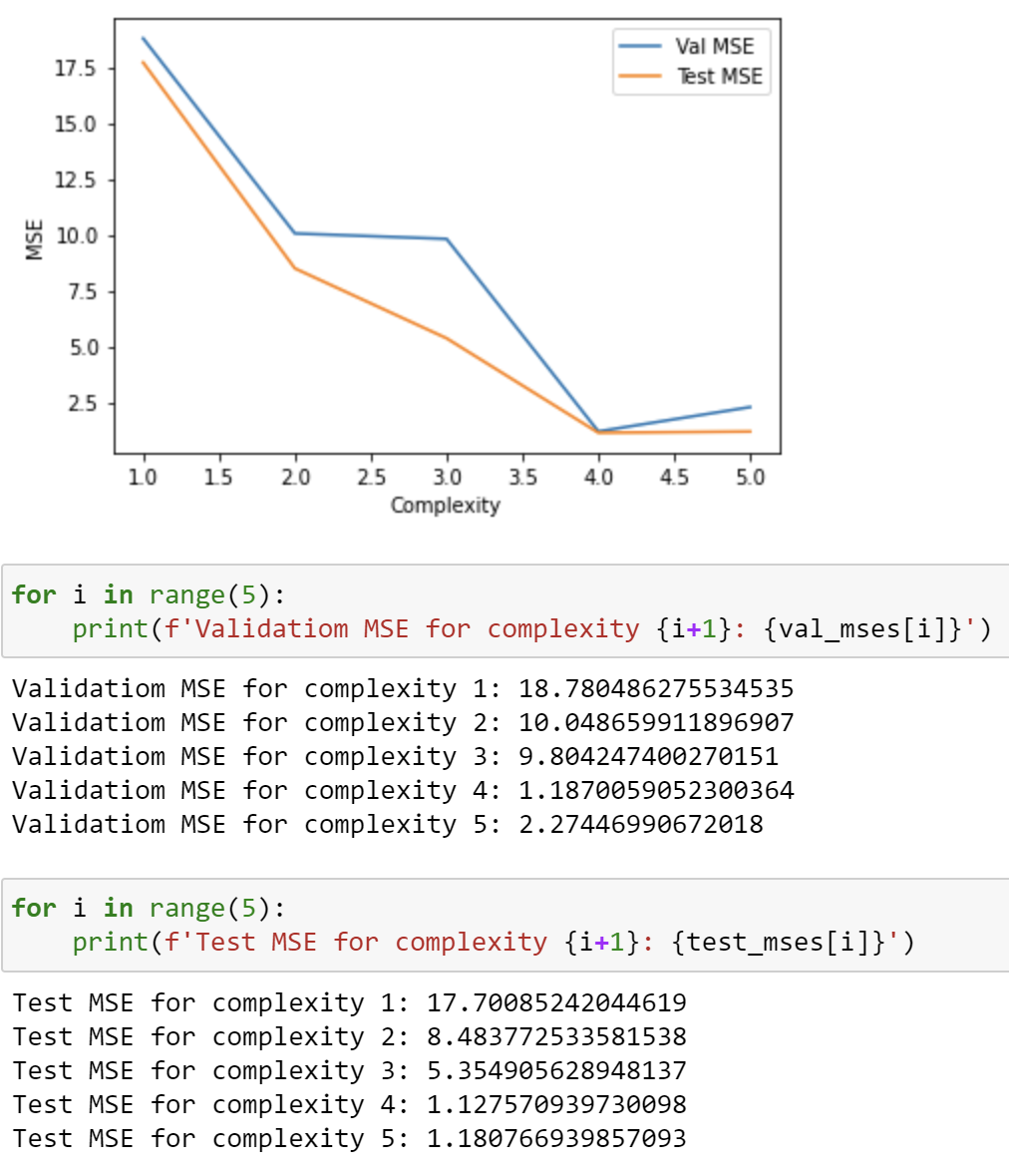
(c)

For seed 5:

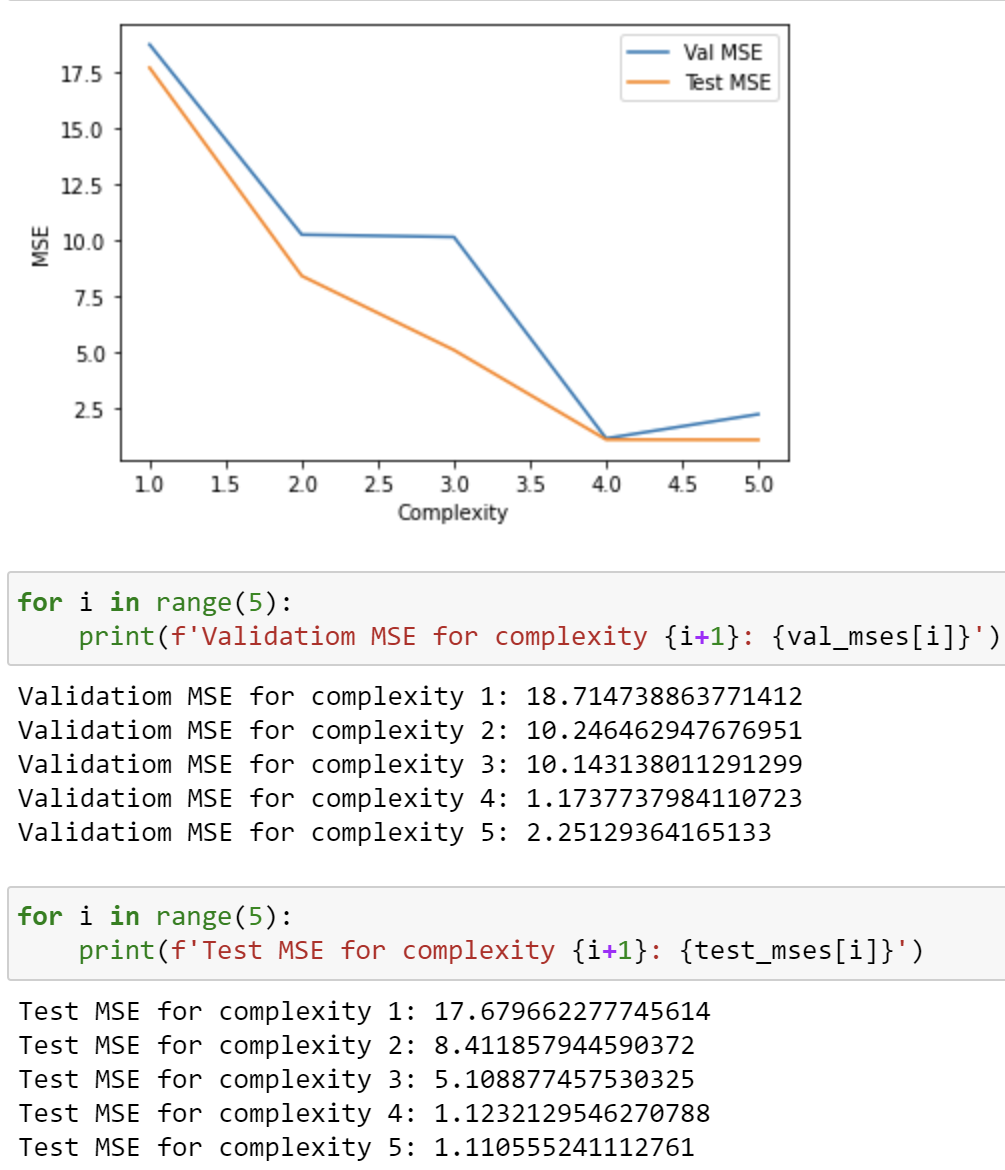
Holdout



Kfold

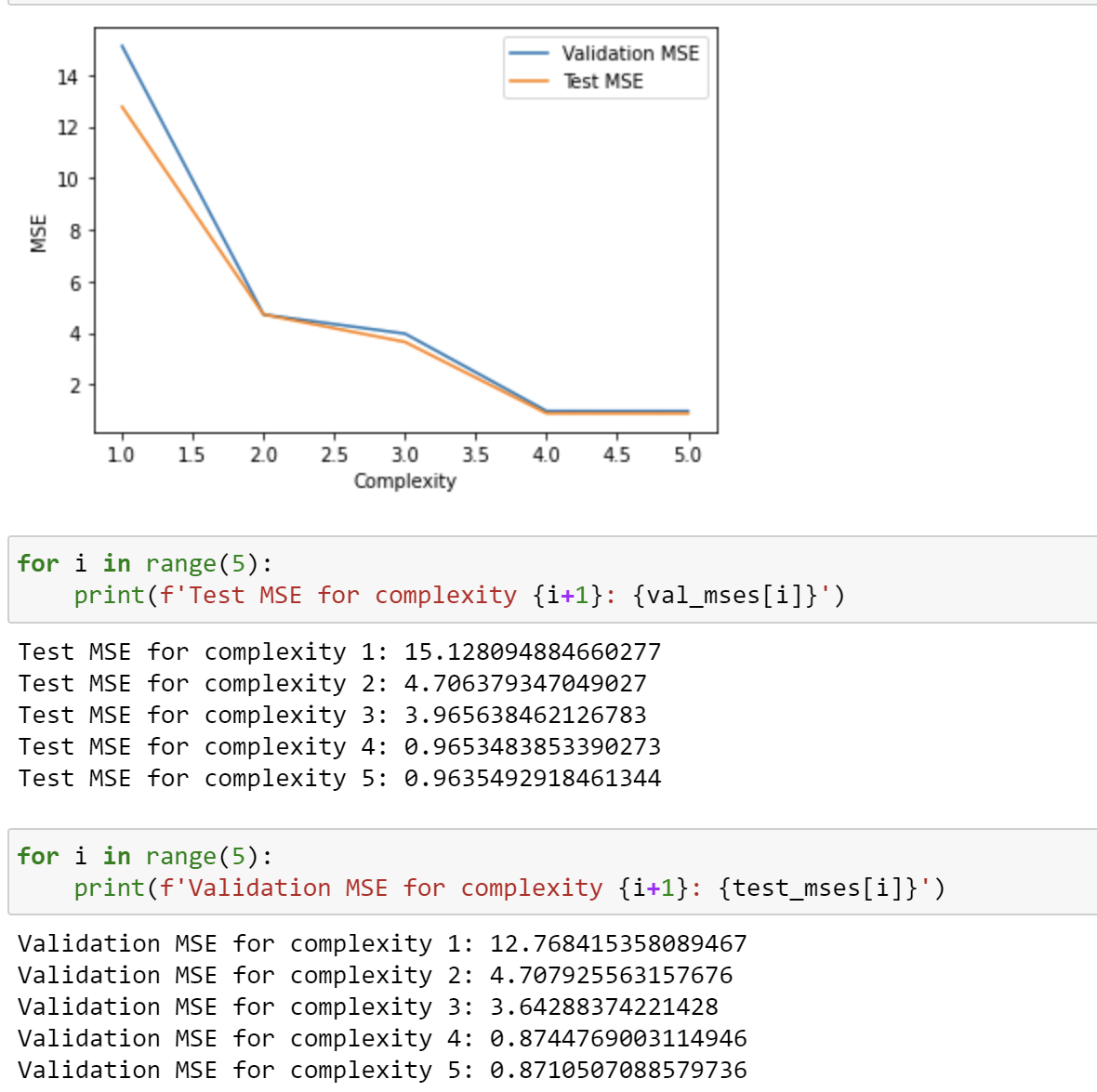


LOOCV

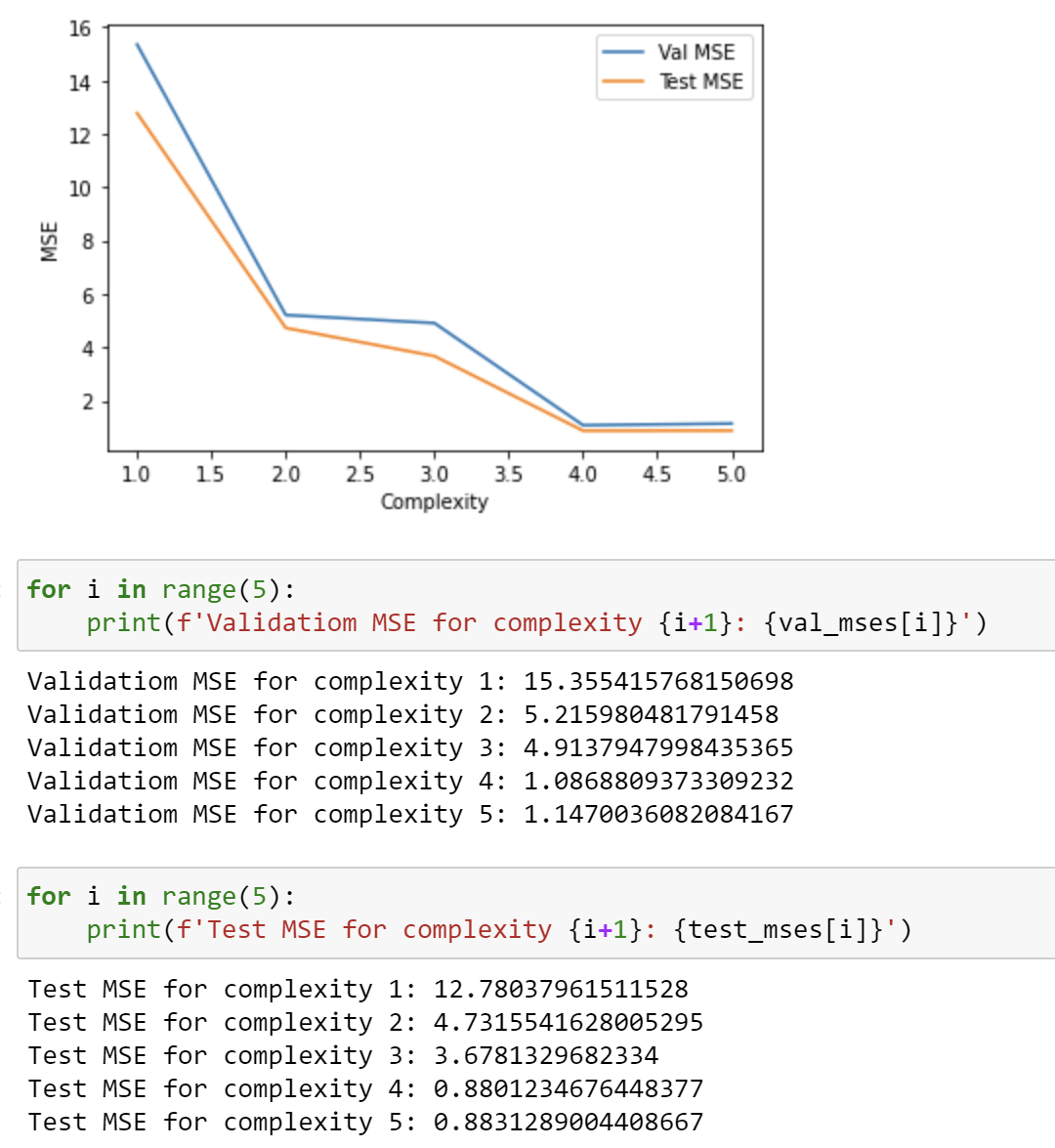


For seed 10:

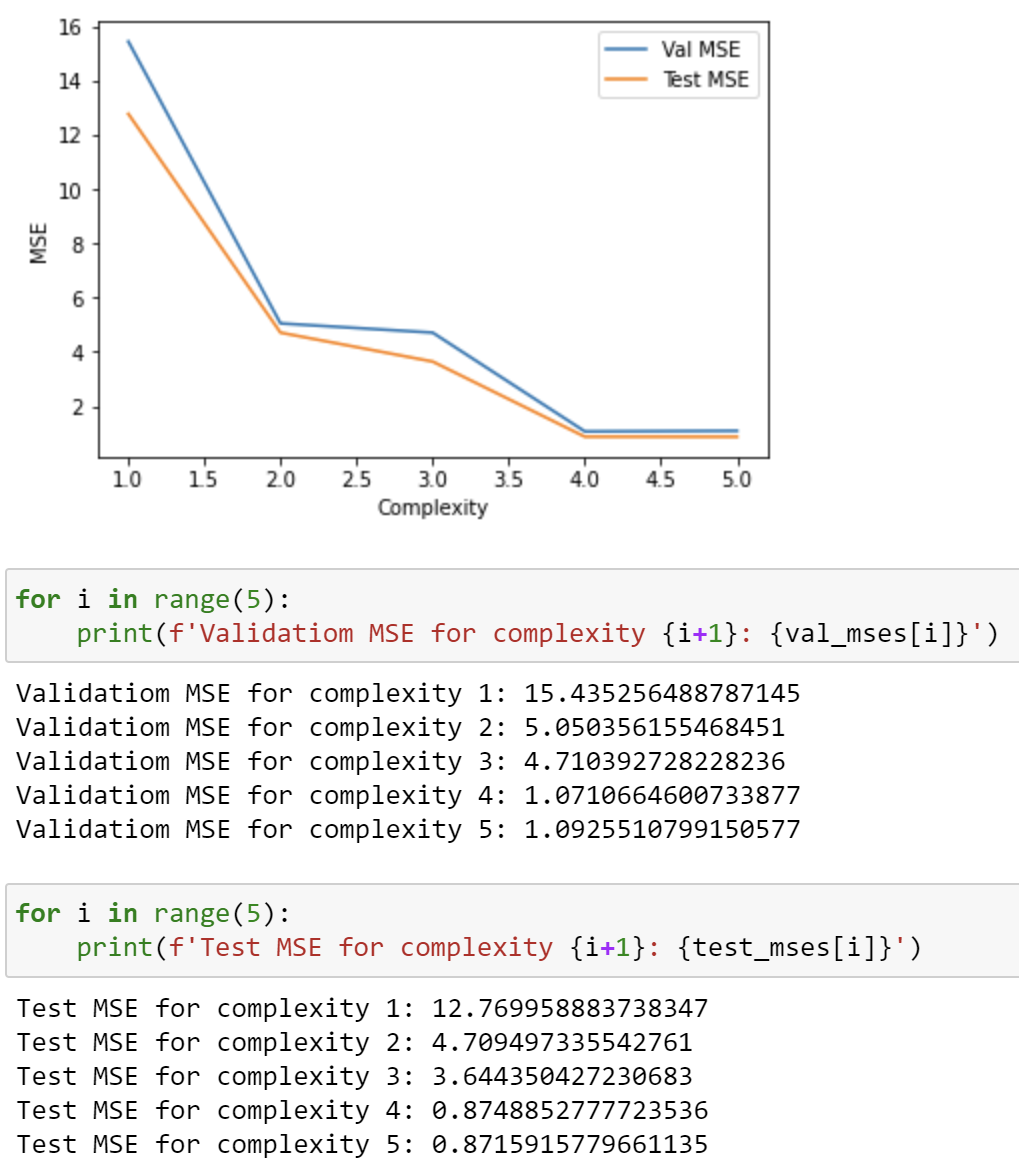
Holdout



Kfold



LOOCV



It is observed that the same trends are preserved across seeds. LOOCV performs best

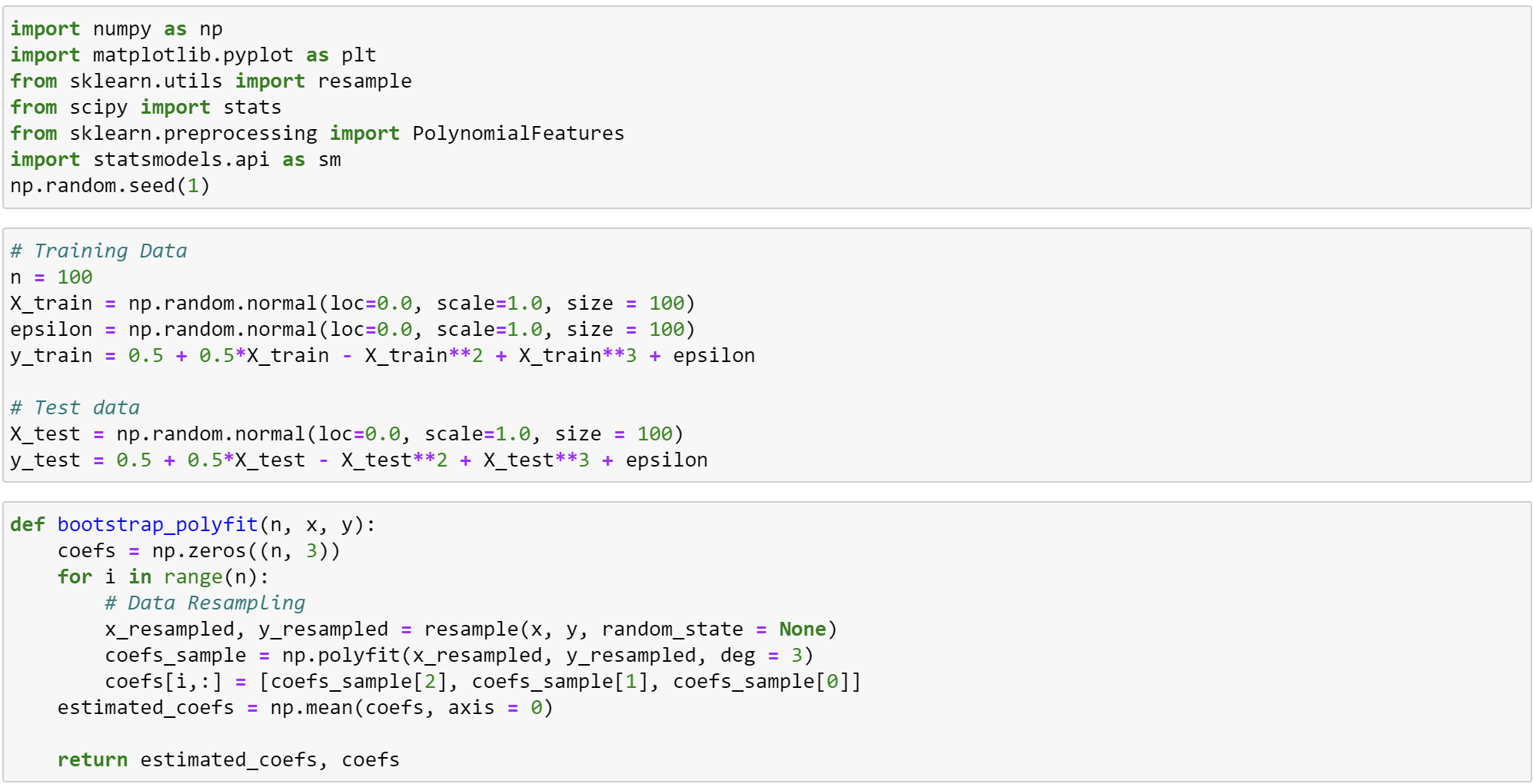
Problem 2:

Problem 2 (24 points): Consider the setup in the previous problem. Our goal is to obtain confidence intervals for parameters of the cubic model by coding up your own bootstrap procedure.

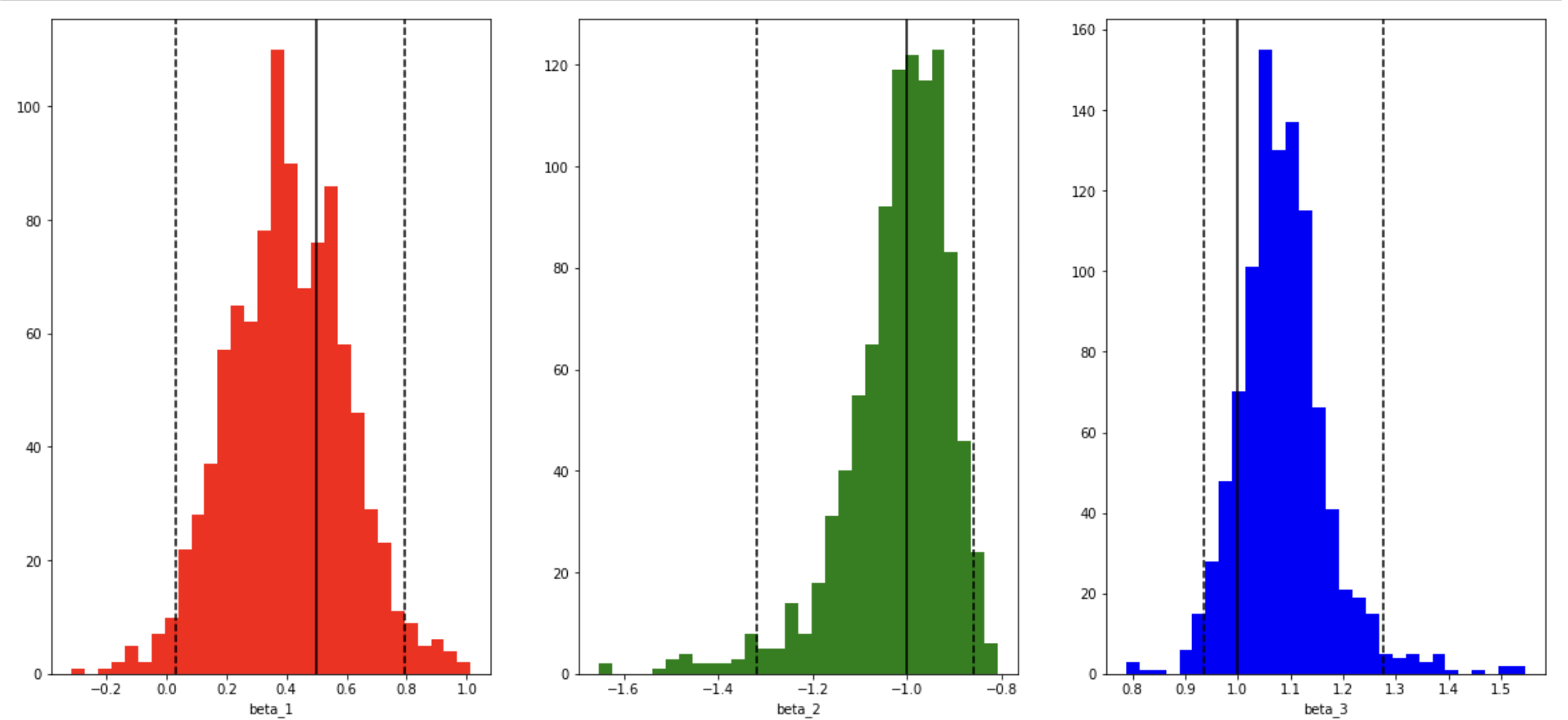
(a) (8 points) Using B = 1000 bootstrap datasets, plot empirical distribution of your coefficients  
for the linear, quadratic, and cubic coefficients.  
(b) (8 points) Based on the previous results, report the 95% interval for each coefficient and  
whether the true coefficient value is in the interval.  
(c) (8 points) Compare your results to the R or python function.

Solution:

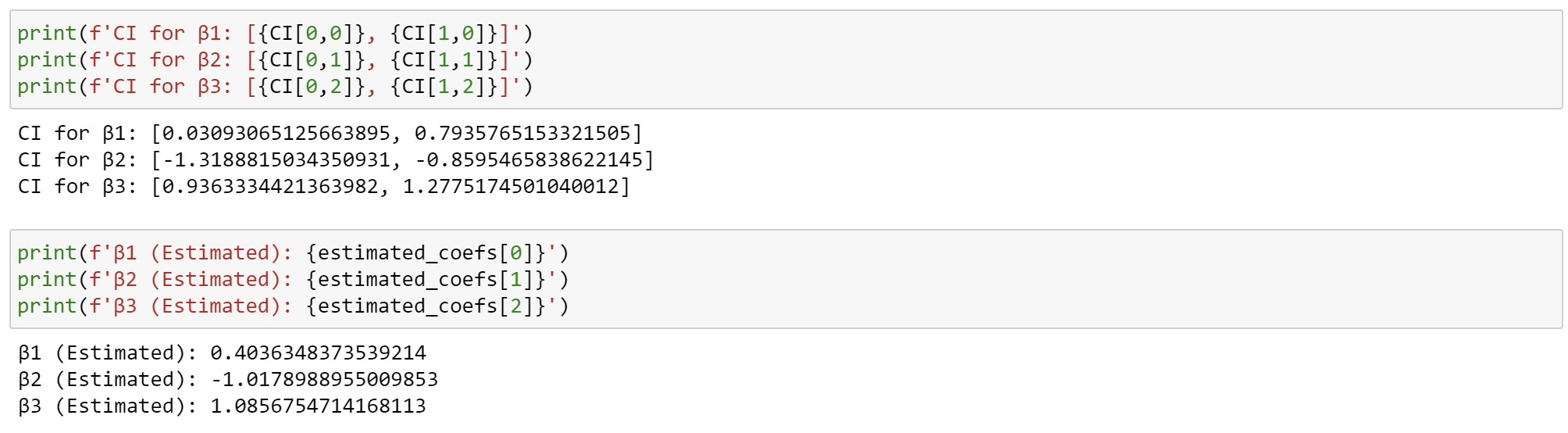
(a)



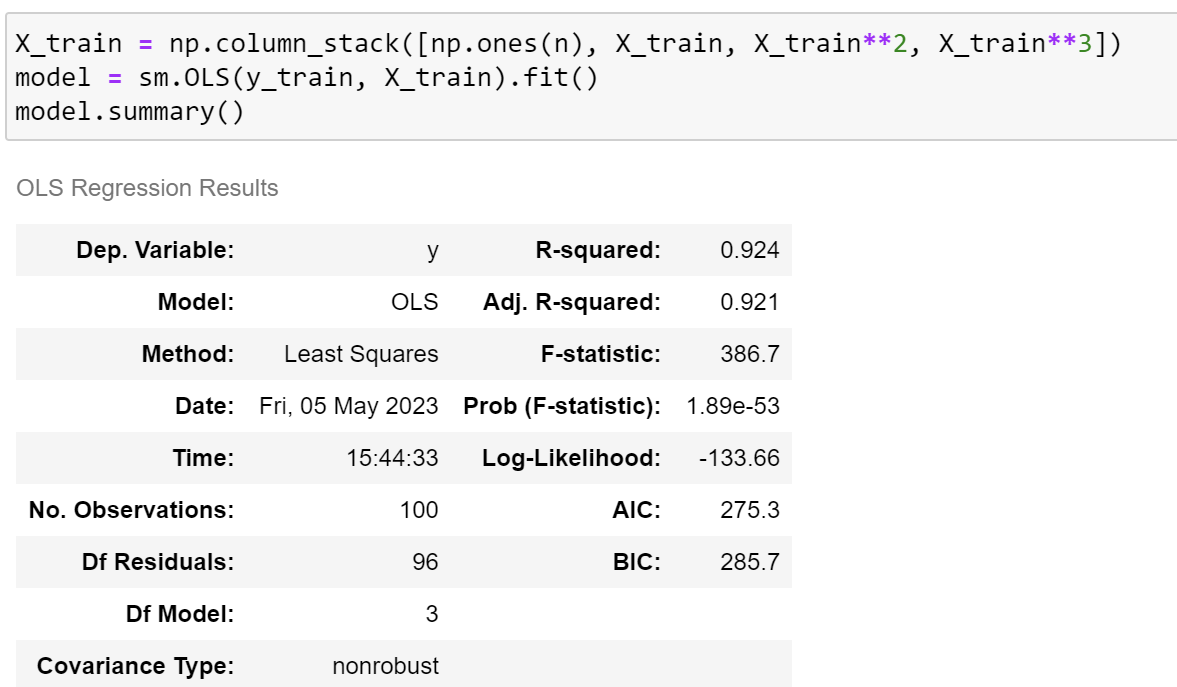


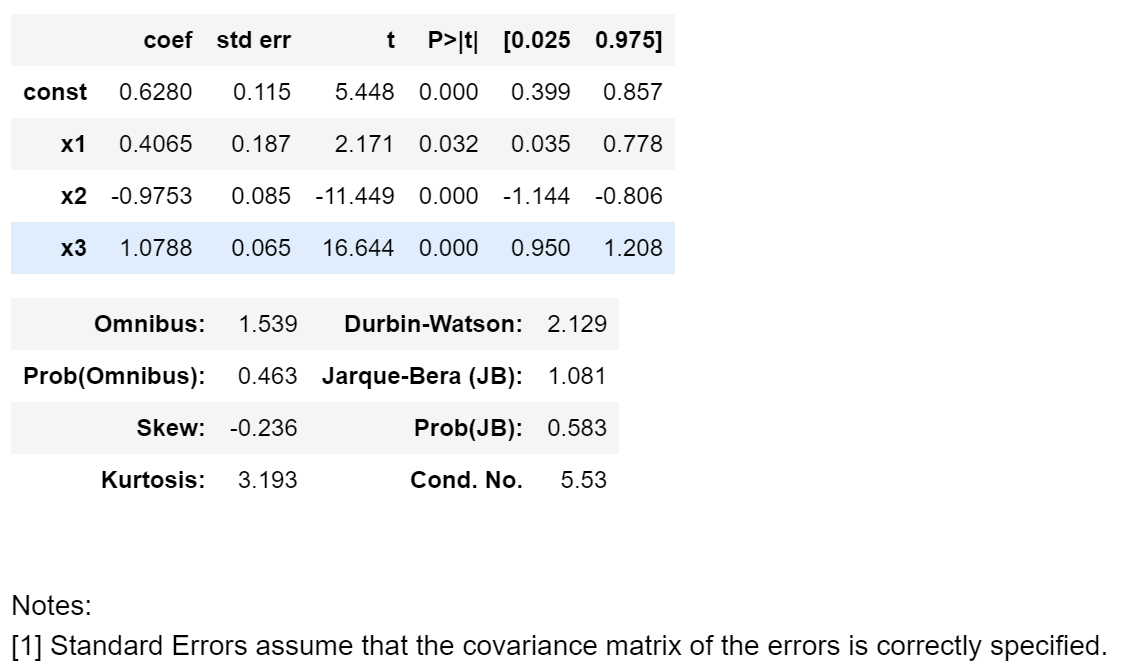


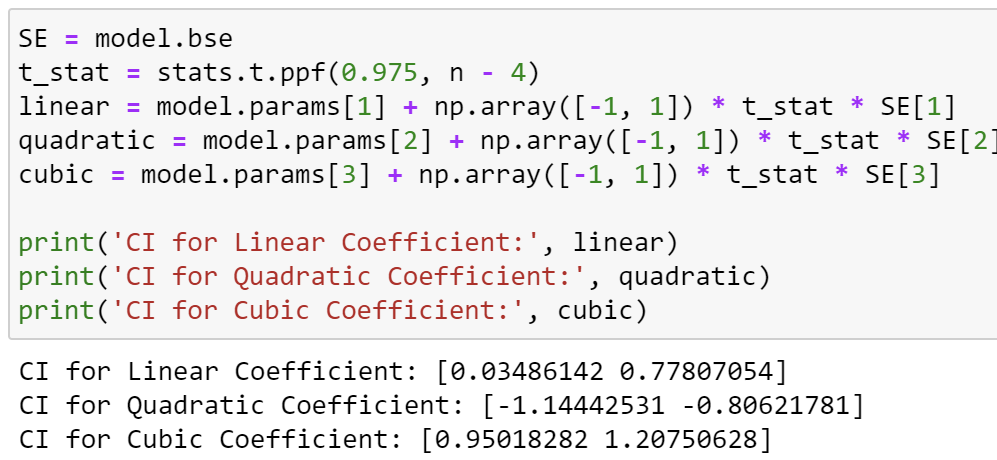
(b)



(c)







Based on the results, although there are minor differences in the absolute values, it appears that the bootstrapping method developed by me is consistent with the results obtained through the statsmodels library.

Problem 4:

The “heart” dataset from the “Heart.csv” file contains information on the presence or absence of heart disease in patients. We will build a logistic regression model to predict the probability of heart disease based on the patient’s age, sex, resting blood pressure, serum cholesterol, etc.

(a) (4 points) Using the glm() function in R (or its correspondence in the statsmodels package in  
Python), estimate the coefficients for age, sex, resting blood pressure, and serum cholesterol  
in the logistic regression model. Compute the estimated standard errors for these coefficients  
using the summary() function.

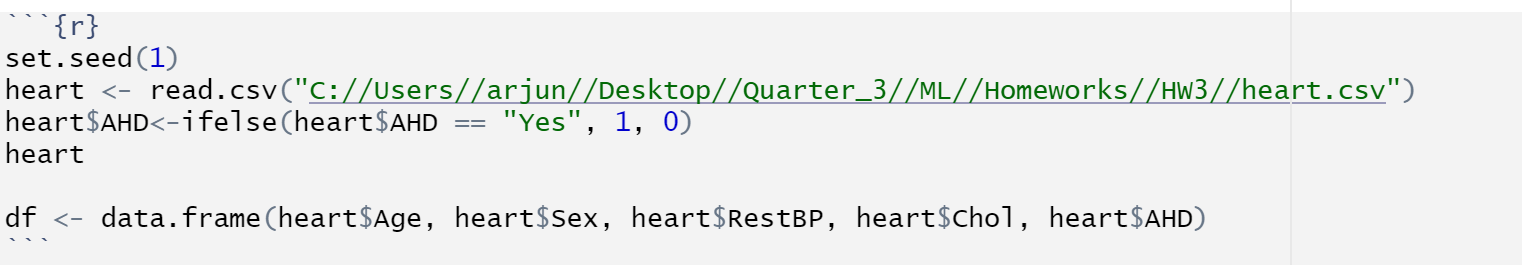
(b) (6 points) Write a function, cv.fn(), that takes as input the dataset and a value for k (the  
number of folds for cross-validation) and outputs the cross-validation estimate of the misclas-  
sification rate for the logistic regression model. You may use the cv.glm() function in R or  
the cross val score() function in scikit-learn (Python) to perform the cross-validation  
(c) (4 points) Use the cv.fn() function to compute the cross-validation estimate of the misclassi-  
fication rate for the logistic regression model.

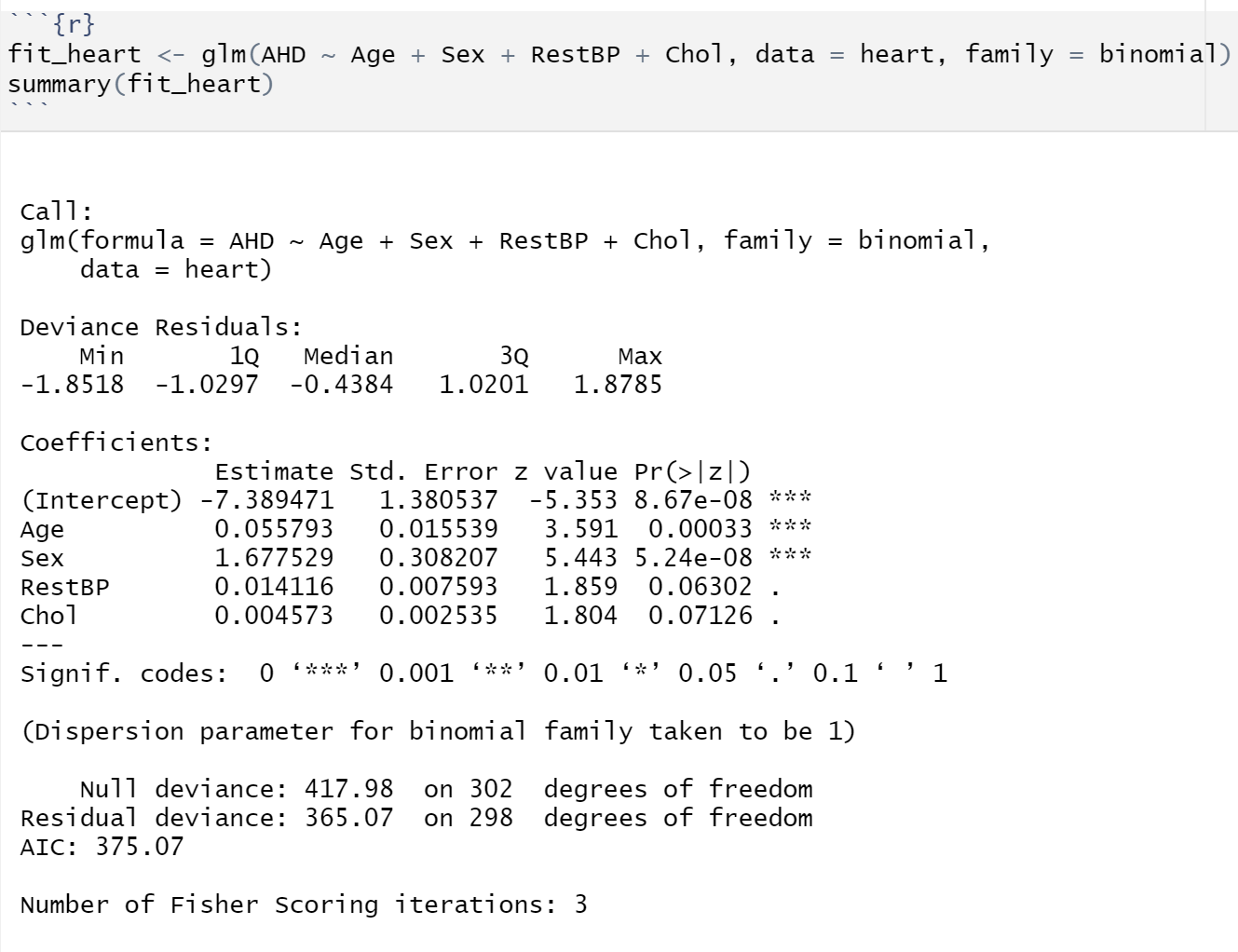
(d) (6 points) Use the bootstrapping technique to perform a bootstrap analysis of the logistic  
regression model. Write a function, boot.fn(), that takes as input the dataset and an index of  
the observations and outputs the coefficient estimates for the logistic regression model. Use  
your boot.fn() function together with the bootstrapping technique to estimate the standard  
errors of the logistic regression coefficients for age, sex, resting blood pressure, and serum  
cholesterol.

(e) Comment on the differences between the estimated standard errors obtained using  
the glm() function and the bootstrap method. Which method do you think is more reliable  
and why?

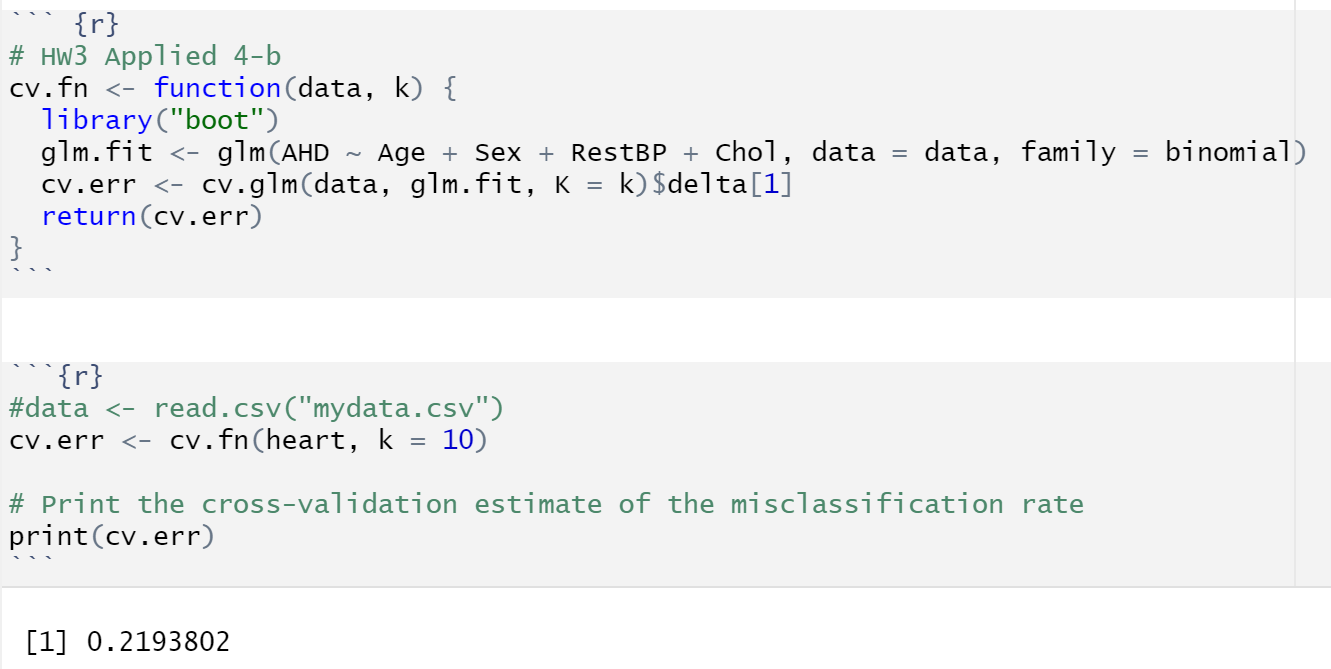
Solution:

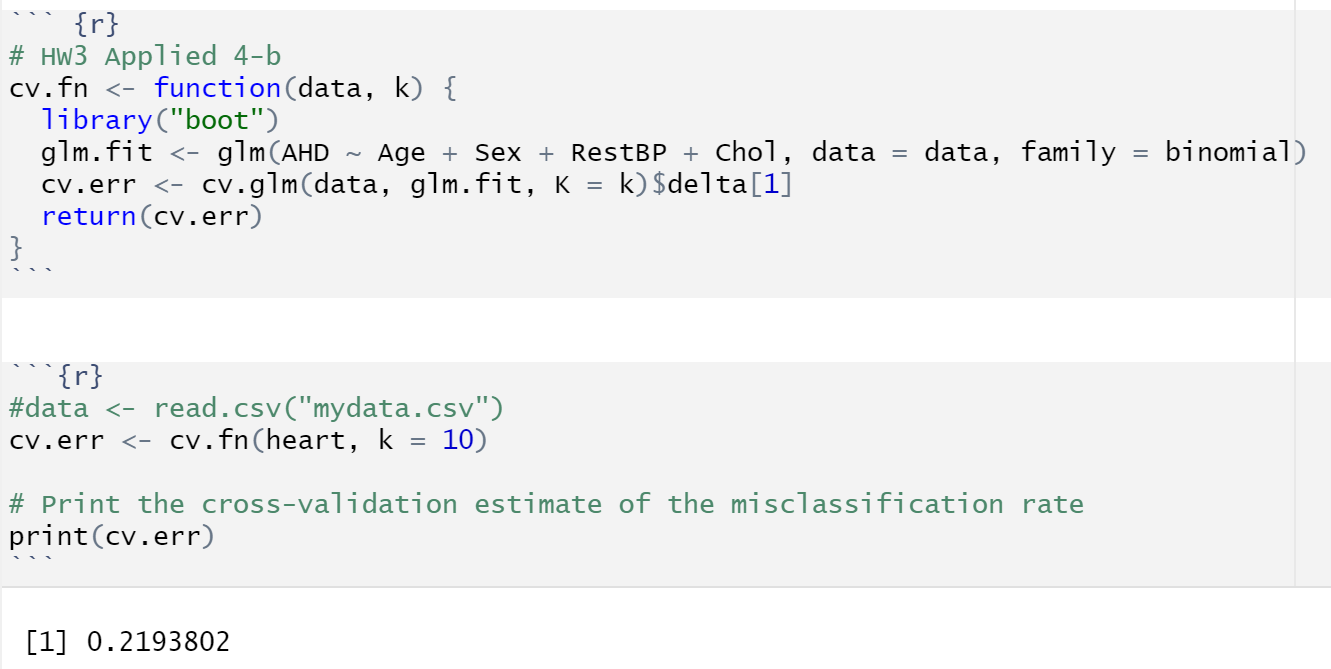
(a)



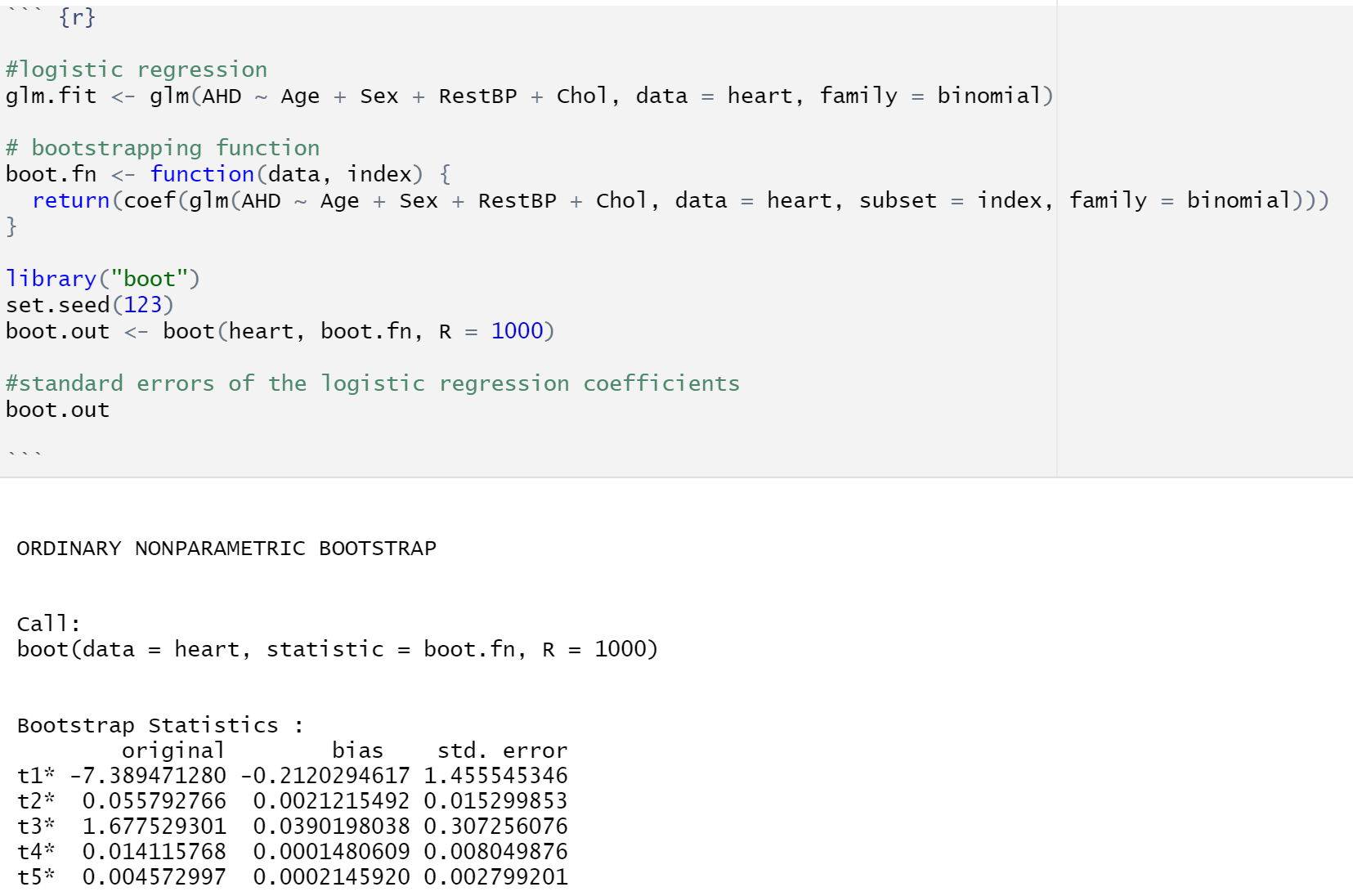


(b)



(c)

(d)



(e)

The difference between traditional Logistic Regression and Bootstrapping is not significant at B = 1000. However, had the bootstrap value been lower, the standard errors would have been significantly higher. Due to the smaller sample size, Logistic Regression appears to be more reliable. Had the sample size been larger, Bootstrapping would have been considered more reliable.

Problem 5:

repeat the previous question but with the opposite programming language. If you used R, now use Python. If you used Python, now use R.