**BU510.650: DATA ANALYTICS, FINAL EXAM – V1**

**SPRING 2020**

Interpreting output of data analysis

Question 1: 10 points

Question 2: 6 points

Question 3: 4 points

Question 4: 7 points

Signing the honor code: 3 points

**TOTAL: 30 points**

Honor code: I have neither given nor received unauthorized aid from a human being, either in person or through other means (e.g., phone, online chat).

**STUDENT NAME: Chengxu Zhou**

**SESSION: 53**

**STUDENT INITIALS (TYPE HERE): CZ**

**BU510.650 – Data Analytics, Spring 2020**

**FINAL EXAM – TAKE HOME**

**Instructor: Diana Prieto**

**Date posted: 03/17/20 at 8:30 am, Due: 03/21/20 at 11:59 pm**

**Exam rules**

- Please submit a word document with your answers and the R file that retrieves your answers

- Late submissions will not accepted. Make sure to allocate time for the submission of your take-home. If you are having issues with Blackboard, please send the exam to my email. Only use my email to submit the exam as a last resource. Please don't submit to my email if you have successfully uploaded your exam through Blackboard.

- The take-home will appear in a blackboard link called "Take home exam" under the project link. The exam submission link will appear shortly but you will not be able to see the exam until March 17 at 8:30 am.

- The take-home is to be completed individually.

- For the take-home you are allowed to consult any type of material that you deem convenient to answer the questions. I am using the word material to define anything that is not people. Consulting textbooks or searching through the web is encouraged.

- The teaching assistants are not allowed to answer questions about the final exam. I reserve the right to answer questions about the final exam. This is because I might consider that answering some questions might create an unfair advantage for the students being helped. Please don't expect answers from me to continue working on the exam. I will respond as soon as I can if I find it proper.

- Unlike the initial plan, the take-home exam will require R coding and all the skills you have developed with the software. Make sure you have all the R material (in-class codes, workshops, homeworks, and project) at hand before the day of the take home. This is my honest advice on how to prepare.

- The take home will require some minimal research. Please expect to be searching some answers that are not explicitly in the material already provided in class. I strongly believe that the exam is doable by any student in the class, as long as you are not under the expectation that everything will be in the R material you already have.

**Exam description**

In this exam, you will use the Cleveland dataset to answer the questions above. At the end of this document, you will find the References and the Appendix. In the References, you will find two webpages. Webpage 1 hosts of the Cleveland dataset, and webpage 2 can be used for consulting commands a cross-checking outputs.

**WARNING: IF AT SOME POINT YOU FIND A COMMAND DEPLOYING AN ERROR THAT CANNOT BE FIXED, PLEASE INVESTIGATE WHY THE ERROR IS OCCURING AND EXPLAIN IT.**

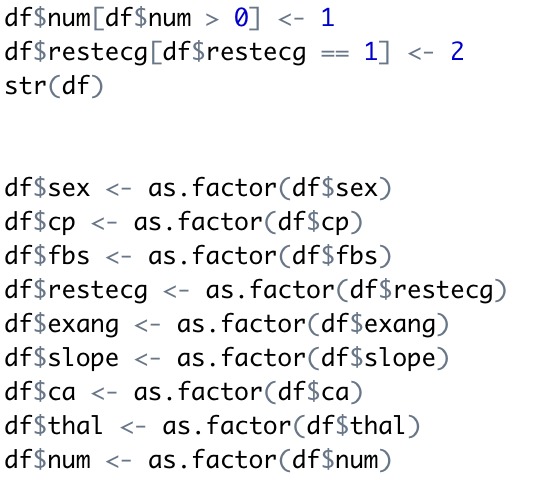
1. **Data preprocessing.**

(2 pts.) Import the Cleveland Dataset

(3 pts.) Delete missing data

(3 pts.) Make sure that all those variables that should be factors are converted into factors, and all the variables that should be numeric are numeric.

(2 pts.) Make sure that the variable “num” is turned into a factor with two categories: 0 and 1 (for categories different from 0).

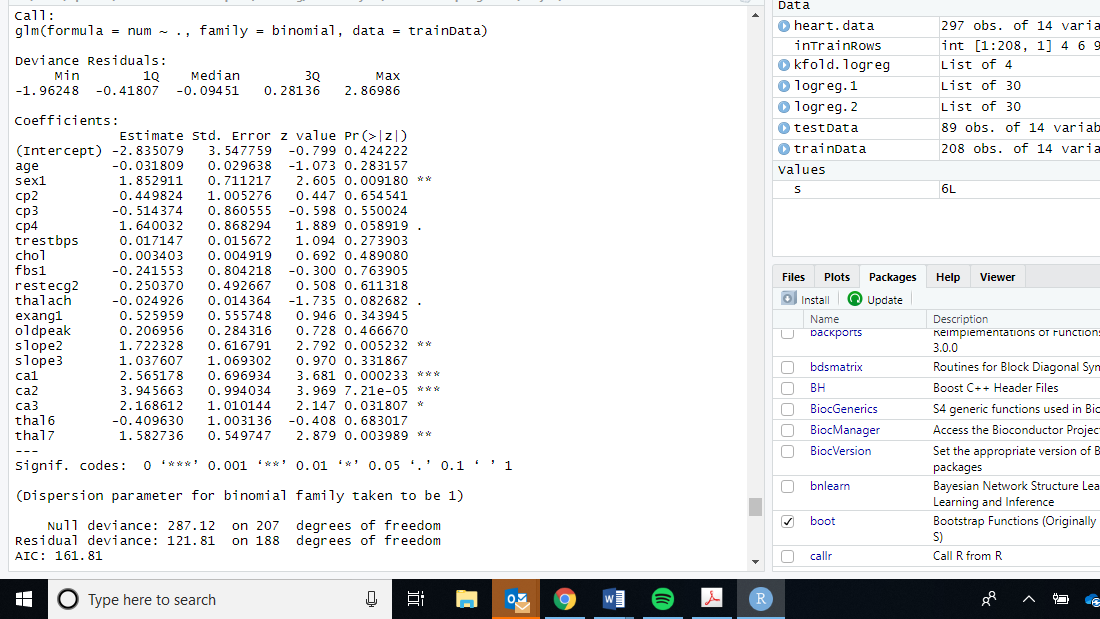
Completed 

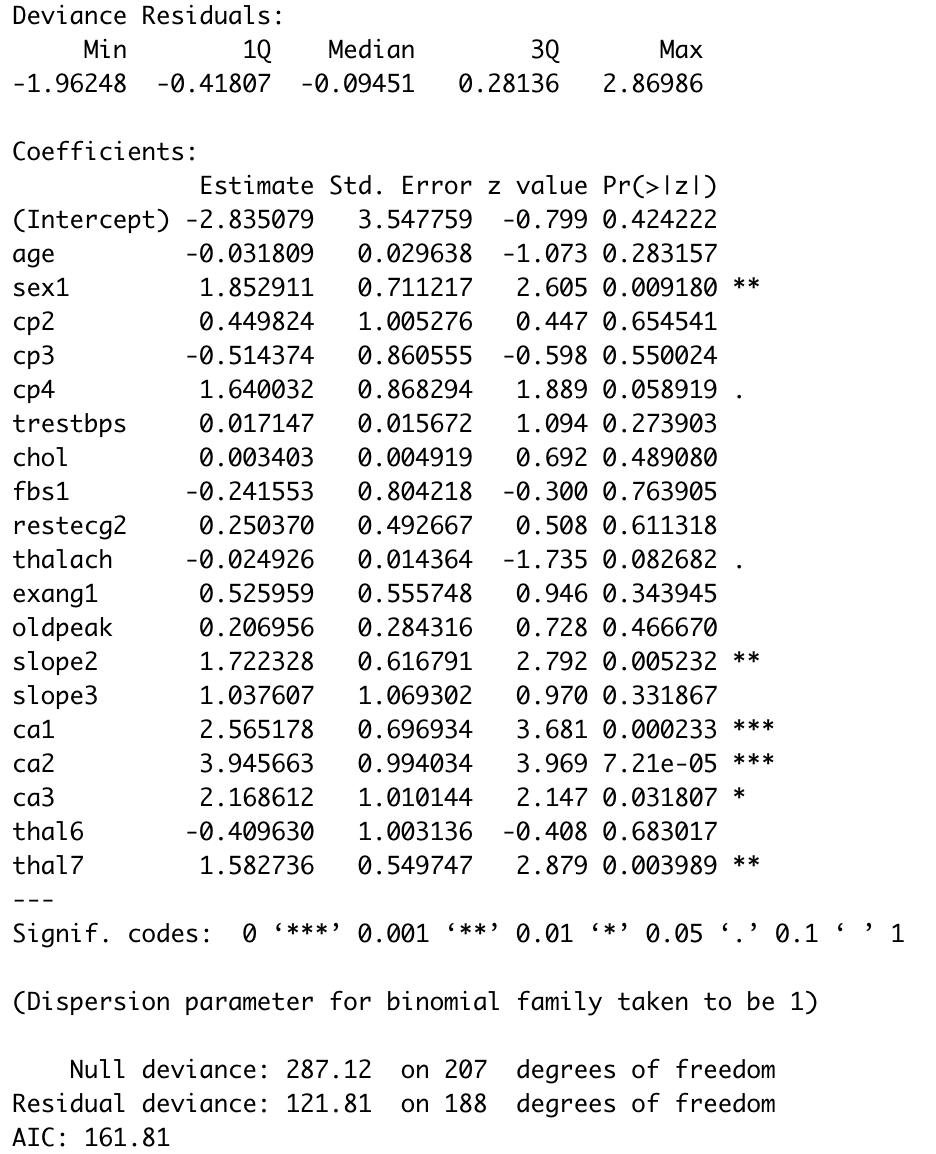
Note: to change restecg1 to 2 is to eliminate singularity.

1. **Model development**

(6 points if you get the exact output and 3 points if you don’t provide the exact output but you provide a good answer)

Randomly split the dataset so that 70% of the data is your training set and 30% of the data is your testing set. Create a logistic regression model that yields the output below. The output should be exact in the parameter estimates and in the p-values.





**I got the same result.**

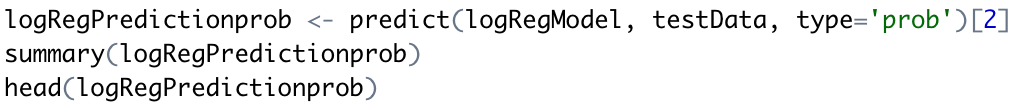
Note that this is the same logistic regression output provided in webpage 2 of the references. You are welcome to use any information on webpage 2 that is relevant to obtaining the output above. Note that doing this step helps validating whether you are on the right track before you try your own coding.

1. **Prediction**

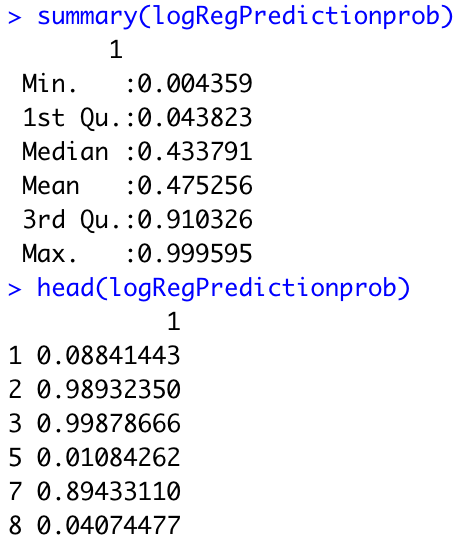
(4 points if you get the exact output and 2 points if you don’t provide the exact output but you provide a good answer)

Predict the probability of heart disease for each of the patients in the testing set developed in the previous point.

Use this code to predict test set:



And the result summary and head are:



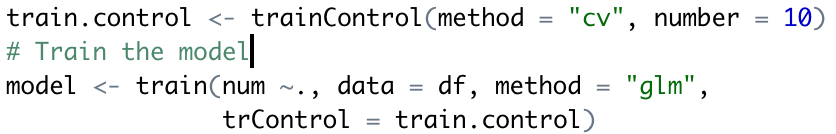
1. **Cross-validation**

(5 pts.) Perform k fold cross validation using a logistic regression model on the complete dataset.

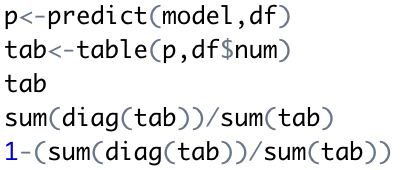
(2 pts.) Retrieve and write the average misclassification error across the k folds.

Hint: You can use the cv.glm function in R

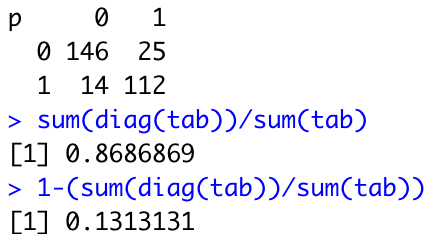
Perform 10 fold validation on full dataset and generate a validated model:



Use that model predict the whole dataset and calculate misclassification error:



The result:



The average misclassification error is 0.131.

1. **Bonus point**

(1 pt) Write a funny joke. We are all in need of good humor!

I drink Corona beer to prevent Coronavirus!

Good luck with your exam!

**References**

**Website 1** <http://archive.ics.uci.edu/ml/datasets/heart+disease>

**Website 2** <https://rpubs.com/mbbrigitte/heartdisease>

**Appendix**

The dataset can be found in the following link:

<https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data>

The description of the columns as they appear in the dataset is as follows:

1. age:  age in years   
   2. sex: sex (1 = male; 0 = female)  
   3. cp: chest pain type   
   -- Value 1: typical angina   
   -- Value 2: atypical angina   
   -- Value 3: non-anginal pain   
   -- Value 4: asymptomatic   
   4. trestbps:  resting blood pressure (in mm Hg on admission to the hospital)  
   5. chol:  serum cholesterol in mg/dl   
   6. fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)  
   7. Restecg: resting electrocardiographic results   
   -- Value 0: normal   
   -- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)   
   -- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria   
   8. thalach: maximum heart rate achieved  
   9. exang:  exercise induced angina (1 = yes; 0 = no)   
   10. oldpeak: ST depression induced by exercise relative to rest  
   11. slope: the slope of the peak exercise ST segment   
   -- Value 1: upsloping   
   -- Value 2: flat   
   -- Value 3: downsloping  
   12. ca: number of major vessels (0-3) colored by flouroscopy  
   13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect  
   14. num: diagnosis of heart disease (angiographic disease status)   
   -- Value 0: < 50% diameter narrowing   
   -- Value 1: > 50% diameter narrowing   
   (in any major vessel)