STAC51: Assignment-2

Shahriar Shams

Fall 2022

Submission deadline: November 28, 2022, 11.59pm (Local Toronto time)

For students registered with AccessAbility, the submission deadline is December 02, 2022, 11.59pm (Local Toronto time)

Late penalty: 10% per day (No submission will be accepted after Dec 04).

Instructions on creating documents for submission

- We will use crowdmark for submission and grading which only accepts PDF, JPG and PNG files.
- I recommend using R-markdown. 1 mark of this assignment is assigned for using R markdown.
- If you do not want to use Rmarkdown, you can write your answers using Microsoft Word and in the end save them as pdfs. But you will lose that 1 mark.
- If you are a Python user, feel free to use Python in place of R to answer any of the questions.
- For each answer, make sure you have provided your codes and outputs. If you are not using R-markdown, take screen shots of your codes and outputs and submit them as part of your answer.
- Make sure your answers are easy to read and nicely presented.

Academic Integrity

Each student will work alone. You are not allowed to ask anyone for help on any platform. Don't ask for solutions to anyone. Do not share your codes or answers. If you need **clarification** on any of these questions, you are allowed to **ask questions on Ed** or ask questions during office hours (please do not email us). And please do not post your solution on Ed and ask "does it look ok?".

When submitting your assignment on crowdmark, there will be a space for an academic integrity statement. Write this following statement on paper/ipad/surface and upload a screenshot of it.

Statement:

I am attesting to the fact that I, [name] (write your full name here), [stnum] (write your student number here), have abided fully to the Code of Behaviour on Academic Matters. I have not committed academic misconduct, and am aware of the penalties that may be imposed if I have committed an academic offence.

Question 1 (7 points)

This question is about fitting a logistic regression model

For this exercise we will use the “Covid-19 cases in Toronto” dataset available on the open data portal of Toronto (https://open.toronto.ca/dataset/covid-19-cases-in-toronto/). [I have filtered the actual data to make sure the models do not have any convergence issue]

Use the “Ever Hospitalized” variable as your response (Y). And use “Age Group”, “Client Gender”, “Outbreak Associated”, “Neighbourhood Name” as your independent variables.

make sure to create a response variable Y which is 1 when Ever Hospitalized=Yes and 0 otherwise

a) Fit a logistic regression model that predicts the event of hospitalization as a function of Age and gender. Interpret all the regression parameter estimates.

b) Check if there is any significant interaction between age and gender. If the interaction is significant, keep the interaction terms in the model.

c) Interpret any one of the interaction terms.

d) List the predicted probabilities for each combination of age and gender.

e) With age and gender already in the model check if Neighbourhood and Outbreak type are needed.

f) Draw the ROC curve for your final model and comment on the AUC.

g) In brief and non technical words, for general audience, explain the findings of your analysis.

Question 2 (4 points)

This question relates to cross validation for logistic regression

We will continue to use the “Covid-19 cases” data set from the previous question.

a) By randomly splitting the dataset in train(60%) and test (40%), comment on the out of sample predictive power of your final model from Question 1.

b) Suppose you are interested in these following two models that predict the event of hospitalization with

   Model-1: covariates: age, gender and interaction between age and gender

   Model-2: covariates: age, gender and interaction between age and gender, neighbourhood

By using a 10-fold cross validation technique, compare the performances of these two models.

Question 3 (4 points)

This question relates to fitting a regression model to multinomial and ordinal data (week-10 materials)

We will use the “Ship” dataset for this task which has three variables in it.

Ship Type = the type of Ship (categories are A, B, C, D and E).

Months of Service = # of months the ship was in service.

Number of Incidents = # of incidents that required repair work.

a) Create a new variable Y that will have three categories

Y = “low” if Number of Incidents ≤ 5

Y = “moderate” if Number of Incidents is > 5 but ≤ 10

Y = “high” if Number of Incidents > 10

Use column “Y” as the response variable.

Suppose you named the dataset “d”. Use this following code to setup the order of variable Y

```
d$Y = factor(d$Y,levels=c("low","moderate","high"),order=T)
```

a) Fit a multinomial regression model with Y as the response, Ship type and months of service as independent variables. Interpret all the regression parameter estimates.

b) Fit an “adjacent category logit” model without proportionality assumption with Y as the response, Ship type and months of service as independent variables. Interpret all the regression parameter estimates.

Code to read the data files in R:

1. Save the csv files in the work folder.

```
#this shows your work folder
getwd()
```

If you are using R markdown, then the folder where you have saved your .rmd file is your work folder.

2. And then use the following line to read any csv file.

```
d= read.csv(file="file name.csv")
```