STOR 665, CWE 2022, Data Analysis

Instructions

General:

- For this open-book exam, the materials you can use are:
  - All course materials, course text, your personal notes, HW problems and any handouts produced during the course are allowed in the exam.
  - Web searching for general queries (e.g. the syntax of particular R or python commands) is allowed.
- Not allowed will be
  - Any form of consultation with another human being during the exam (except that you are allowed to send queries to the instructors).
  - Searches related to the specific datasets used in the exam. You are not allowed to try to search for that publication or anything that might be specific to the dataset used.
- You have **four** hours for this open-book exam. Please submit your exam before **5pm** on Gradescope or email it to the instructors.

STOR 665 instruction:

- Please find the exam paper and code templates at the following link: : `https://drive.google.com/drive/folders/1xsbrJ8BV4_0TQwK91TFn6mK8Xmc2y68h?usp=sharing`. If you have any problem getting access to the exam materials, please contact the instructor (yaoli@email.unc.edu) for help immediately.

- Please use the R markdown and Jupyter notebook templates to answer the two data analysis questions. Include your **code**, **code output** and **answers** in the pdf output of R markdown and Jupyter notebook.
- Submission: Please submit the pdf outputs of each question to the Gradescope site. If you have problem submitting with Gradescope, you can email the submission to the instructor (yaoli@email.unc.edu).
- If you have any questions during the exam, you can email the instructor (yaoli@email.unc.edu) for help.

STOR665

1. The bioChemist dataset includes number of articles produced by graduate students in biochemistry during last 3 years of Ph.D. programs (ART), gender (GEN), marital status (MAR), number of kids (KID), prestige of Ph.D. department (PRE), and number of articles produced by Ph.D. mentor during last 3 years (MENT). The data set was generated to study the relationship between productivity of biochemistry Ph.D. students and related variables.

(a) Analyze the data with a Poisson regression model (canonical link). Do the predictors have effects on the response?

(b) For the two slope parameters of KID and MENT construct the simultaneous 95% confidence ellipsoid for them. Use the simultaneous confidence ellipsoid construction to determine a p-value for the null hypothesis that both slope parameters are 0 (Note: You can use R package *ellipse*).

(c) Investigate whether there are interactions between the predictors. Write the overall null hypothesis for no interactions in the form $H_0 : \mathbf{A}\beta = 0$ and determine the df. Select the interactions which you believe should be included in the final model. Provide your reasons and interpret your model, taking into account both main effects and interactions.

(d) Is there evidence for zero-inflation? Provide your conclusion and reasons.

2. The CIFAR10 dataset contains natural images of 10 different classes. In this exercise, we only use the first two class, airplane and auto, for binary classification. Please implement one model from scratch to do the binary classification (airplane vs. auto). Code of data loading can be found in the code template.

(a) Describe your model. (If DNN is used, you can print the model architecture in python as the answer.)

(b) Train the model on the training set and report the average batch loss for each epoch.

(c) Report the training parameters: number of epochs, optimizer and related parameters (learning rate, penalty, momentum etc).

(d) Test the model on the test set and report the testing accuracy.