ACTSC 632 – Spring 2023
Test 2 – July 13, 2023 – 10am-11:50am

Instructions:

- You will be given a total of 110 minutes to write this examination. At the end of the examination, you will be asked to upload your solutions on the **Test 2 dropbox in Learn** (I will provide extra time for ONLY the upload)**.**

- This examination will take place in M3 2134 (our usual classroom for the course) on **Thursday July 13, 2023** from **10:00am to 11:50am**. You will be given access to the exam paper and the data set to be analyzed on Learn (in the folder Tests under Content) 5 minutes before the start time of the examination (i.e., at 9:55am).

- You are expected to complete this examination on your own personal laptop/device. Please ensure that this laptop/device is one that you are familiar to work with and on which a recent version of R or RStudio is installed. (You may find it useful to have installed the library "dplyr", which has been used many times in Module 2.) If there are any issues with this directive, please let me know as soon as you can.

- This is an open-book examination. However, you must complete this examination solely on your own. **You are strictly forbidden to discuss or receive any external/third-party help to complete this examination.**

- Once you have downloaded the exam paper and dataset, you will be asked to turn off the wifi on your laptop until the end of the exam. At 11:50am, when the exam ends, you will be given time to upload your answers in the Learn dropbox.

- As this is an open-book exam but where you will not have internet access during the exam, you are encouraged to download on your personal laptop/device any course materials you may want to refer to during the exam.

- The exam paper has only 1 long question, with many sub-questions.

- Checklist of documentation to be submitted on the Test 2 dropbox in Learn:

  - a pdf (or html) document including your answers to each of the sub-questions (properly annotated for ease of grading);
  - if not included in the above document, the actual R code you develop to complete this examination;
  - any additional documentation you judge relevant.

- Please organize the documentation submitted so that it is easy for someone else to navigate it and understand your thought process in reaching your conclusions. Clarity of thought and work tends to increase the partial credits you may receive.

  Good Luck!

Question [50pts] You are asked to conduct a tariff analysis on the data set `KenyaCarInsurance` (see csv file in the folder Tests on Learn) pertaining to the claim experience of a group of automobile insurance policies over the years 2015 to 2018 in Kenya. First download the file and use the command `read.csv` in R to read it.

Then, familiarize yourself with the data set. Note that each row of the data set has the following claim history information aggregated among all policies with a given set of rating factors:

- `exposure_years`: total number of policy years of experience collected within each tariff cell, also referred to as "duration" in class and below;

- `nb_claims`: total number of claims recorded for each tariff cell;

- `severity`: average claim size recorded for each tariff cell

The data set also includes information about the earned premium income for each tariff cell, and the total cost per cell (from which the severity was computed).

In addition, the column "Category" provides you with information about the encoding used for the different rating factors.

1. [2pts] By examining the `KenyaCarInsurance` data set, state how many tariff cells there are. Also, determine how many rating factors we have, as well as the number of categories for each rating factor.

2. [3pts] You are asked to set as the base tariff cell (for both the frequency and severity models) the one with the largest duration. Identify explicitly this base tariff cell.

3. [4pts] Model the frequency data using a glm based on the relative Poisson distribution and a log-link function. Using the concept of deviance, discuss the overall fit of this model. Does the relative Poisson glm model seem reasonable to fit the frequency data? Comment.

4. [4pts] Now model the frequency data using a glm based on the relative quasi-Poisson distribution, still with a log-link function. Using the concept of deviance, discuss the overall fit of this model. Does the relative quasi-Poisson glm model seem reasonable to fit the frequency data? Comment.

5. [6pts] Propose a potential simplification to the relative quasi-Poisson glm model of Q4 by dropping one of the rating factors and explain your choice. Via a likelihood ratio test argument, determine whether or not you are statistically justified to simplify the relative quasi-Poisson glm model of Q4. State the null and alternative hypothesis of this likelihood ratio test together with its test statistics. Continue the test with the model chosen in this sub-question.

6. [7pts] Using the relative quasi-Poisson glm model chosen in Q5:

   (a) [3pts] estimate the multipliers (i.e., relativities) for the frequency data and provide a 95% confidence interval for each.

   (b) [4pts] identify the tariff cell with the largest expected number of claims per exposure period. Find the expected number of claims per year for a policy in this tariff cell.

7. [4pts] Model the severity data using a glm based on the gamma distribution and a log-link function. Using the concept of deviance, discuss the overall fit of the model. Does the gamma glm model seem reasonable to model the claim severity data set? Comment.

8. [6pts] You are now asked to combine the age category 6 (70 and above) with the age category 4 (31-50) in the gamma glm model of Q7. Using a likelihood ratio test, comment on the appropriateness to do so. State the null and alternative hypothesis of this likelihood ratio test. Are you statistically justified to simplify the model in Q7 to the one suggested here? Comment.

9. [9pts] Based on your recommendation in Q8:

   (a) [3pts] estimate the multipliers (i.e., relativities) for the severity data and provide a 95% confidence interval for each.

   (b) [3pts] identify the tariff cell with the largest expected claim size. Find this expected claim size.

   (c) [3pts] identify the tariff cell with the smallest expected claim size. Find this expected claim size.

10. [5pts] Based on your multiplier estimates for the frequency (obtained in Q6) and severity data (obtained in Q9), find the pure premium for a set of 60 independent 1-year policies of which a third are in tariff cell `age = 2, gender = 2, license = 2` while the remaining are in tariff cell `age = 6, gender = 1`, and `license = 1`.