MSCA31010: Linear & Non-Linear Models

Winter 2021 Assignment 3

# Questions 1 and 2

The **Homeowner\_Claim\_History.xlsx** contains the claim history of 27,513 homeowner policies. The following table describes the eleven columns in the HOCLAIMDATA sheet.

| **Name** | **Description** | **Categories** |
| --- | --- | --- |
| policy | Policy Identifier |  |
| exposure | Number of Exposure Units |  |
| num\_claims | Number of Claims |  |
| amt\_claims | Total Amount of Claims |  |
| f\_primary\_age\_tier | Age Tier of Primary Insured | < 21, 21 - 27, 28 - 37, 38 - 60, > 60 |
| f\_primary\_gender | Gender of Primary Insured | Female, Male |
| f\_marital | Marital Status of Primary Insured | Not Married, Married, Un-Married |
| f\_residence\_location | Location of Residence Property | Urban, Suburban, Rural |
| f\_fire\_alarm\_type | Fire Alarm Type | None, Standalone, Alarm Service |
| f\_mile\_fire\_station | Distance to Nearest Fire Station | < 1 mile, 1 - 5 miles, 6 - 10 miles, > 10 miles |
| f\_aoi\_tier | Amount of Insurance Tier | < 100K, 100K - 350K, 351K - 600K, 601K - 1M, > 1M |

# Question 2 (25 points)

Using all the seven categorical predictors *f\_primary\_age\_tier*, *f\_primary\_gender*, *f\_marital*, *f\_residence\_location*, *f\_fire\_alarm\_type*, *f\_mile\_fire\_station*, and *f\_aoi\_tier* to define segments of observations, please estimate the Tweedie’s parameter and the dispersion parameter.

# Question 3 (75 points)

Train a Pure Premium model to predict the total claim amount with the following specifications.

1. The target variable is *amt\_claims*
2. The predictors are *f\_primary\_age\_tier*, *f\_primary\_gender*, *f\_marital*, *f\_residence\_location*, *f\_fire\_alarm\_type*, *f\_mile\_fire\_station*, and *f\_aoi\_tier*
3. The distribution assumption is Tweedie with your parameter in Question 2
4. The link function is the logarithm
5. The offset variable is the logarithm of *exposure*
6. The model must include the Intercept term
7. Use the Forward Selection method to enter significant predictors
8. Use the Deviance statistic to select the predictor entered in each step
9. The tolerance level is 5%
10. (40 points). Show the Forward Selection summary table. The table should contain only the predictors that are selected to enter. The columns are Predictor’s Name, Number of Non-aliased Parameters, Quasi-Log-Likelihood, Deviance Chi-Square, Deviance Degree of Freedom, and Deviance Significance.
11. (15 points). Show the complete set of parameter estimates (including the aliased parameters). Please also include the exponentiated parameter estimates.
12. (15 points). Plot the predicted claim amount versus the observed claim amount. Please use the exposure to represent the color of the markers. A gradient color bar should be included.
13. (5 points). Please comment on the predictions based on the scatterplot in (c).

# Optional Question 2 (20 points)

What is the canonical link function of the Tweedie distribution? **Hint**: If we express the Tweedie density function as and , then and . Suppose is the canonical link function, then . Your answer must match the canonical link function for the known distributions for Normal, for Poisson, for Gamma, and for Inverse Gaussian. Please show your works.