MSCA31010: Linear & Non-Linear Models

Winter 2022 Assignment 2

In insurance ratemaking, the term Frequency is defined as the number of claims divided by the duration of exposure of a policy. Using the claim\_history.csv, we will study how policy attributes affect Frequency. You are asked to train a generalized linear model with the following specifications.

* Response Variable: CLM\_COUNT
* Offset Variable: Natural logarithm of EXPOSURE.
* Categorical Predictors: MSTATUS, CAR\_TYPE, HOMEKIDS, KIDSDRIV, REVOKED, and URBANICITY (Please reorder the categories of each categorical predictor in ascending order of number of observations)
* Interval Predictors: CAR\_AGE, MVR\_PTS, TIF, and TRAVTIME
* Distribution: Poisson
* Link Function: Natural logarithm

Your goal is to train a model that has just the right set of predictors. You must perform the calculations without calling any special libraries (e.g., scikit-learn or statsmodels). The standard libraries such as matplotlib, numpy, pandas, and scipy are allowed. You need to drop all missing values (i.e., NaN) of all the predictors, the target variable, and the offset variable before training your model.

# Question 1 (20 points)

Before you train the model, you want to explore the predictors.

1. (20 points) For each predictor, generate a scatterplot chart that shows the number of claims by the predictor’s values. Also, color-code the markers by the exposure values. Please display the predictor’s values are displayed in ascending lexical order.

# Question 2 (40 points)

Enter the predictors into your model using Forward Selection. The Entry Threshold is 0.05.

1. (15 points). Please provide a summary report of the Forward Selection. The report should include (1) the step number, (2) the predictor entered, (3) the number of non-aliased parameters in the current model, (4) the log-likelihood value of the current model, (5) the Deviance Chi-squares statistic between the current and the previous models, (6) the corresponding Deviance Degree of Freedom, and (7) the corresponding Chi-square significance.
2. (5 points). What predictors does your final model contain?
3. (5 points). What are the aliased parameters in your final model? Please list the predictor’s name and the aliased categories.
4. (5 points). How many non-aliased parameters are in your final model?
5. (10 points). Please show a table of the complete set of parameters of your final model (including the aliased parameters). Besides the parameter estimates, please also include the standard errors, and the 95% asymptotic confidence intervals. Conventionally, aliased parameters have missing standard errors and confidence intervals.

# Question 3 (20 points)

You will visually assess your final model in Question 2. Please color-code the markers according to the Exposure value.

1. (10 points). Please plot the predicted number of claims versus the observed number of claims.
2. (10 points). Please plot the Deviance residuals versus the observed number of claims.

# Question 4 (20 points)

You will calculate the Accuracy metric to assess your final model in Question 2.

1. (10 points). Please calculate the Root Mean Squared Error, the Relative Error, and the R-squared metrics.
2. (10 points). Please comment on the Final Model based on the above three metrics and the diagnostic charts in Question 3.