MSCA31010: Linear & Non-Linear Models

Winter 2021 Assignment 2

You are asked to train a binary logistic regression model on the claim_history.csv. Your model will predict the likelihood of filing more than one claim in one unit of exposure. You will first calculate the Frequency variable by dividing the CLM_COUNT by EXPOSURE. Next, you will create a binary target variable that determines if the Frequency is strictly greater than one (i.e., the Event).

You will use MSTATUS, CAR_TYPE, REVOKED, and URBANICITY as the categorical predictors, and CAR_AGE, MVR_PTS, TIF, and TRAVTIME as the interval predictors. 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 numpy and pandas are allowed. You need to drop all missing values (i.e., NaN) of all the predictors and the target variable before training your model.

Question 1 (25 points)

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

a) (15 points) For each predictor, generate a line chart that shows the odds of the Event by the predictor's unique values. The predictor's unique values are displayed in ascending lexical order.
b) (10 points) Also, calculate the ratio of the maximum odds value to the minimum odds value. If the minimum odds value is zero, then the ratio is infinity. Based on the ratio, please provide us your opinions of whether the final model will include that predictor.

Question 2 (40 points)

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

a) (20 points). Please provide a detailed report of the Forward Selection. However, you do not need to show steps such as 1.1. The report should include (1) the predictor entered, (2) the number of free parameters, (3) the log-likelihood value, (4) the Deviance Chi-squares statistic, (5) the Deviance Degree of Freedom, and (6) the Chi-square significance.
b) (5 points). Which predictors does your final model contain?
c) (5 points). What are the aliased parameters in your final model? Please list the predictor's name and the aliased categories.
d) (5 points). How many non-aliased parameters are in your final model?

e) (5 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 exponentiated estimates (i.e., apply the exp() function on the parameter estimates).

Question 3 (20 points)

You will visually assess your final model in Question 2. Please color-code the markers according to the Exposure value. Also, please briefly comment on the graphs.

a) (10 points). Please plot the predicted Event probability versus the observed Frequency.

b) (10 points). Please plot the Deviance residuals versus the observed Frequency.

Question 4 (15 points)

You will calculate the Accuracy metric to assess your final model in Question 2. If the predicted Event probability of an observation is greater than or equal to 0.25, then you will classify that observation as the Event (i.e., filing more than one claim per unit exposure). An observation is correctly classified if the predicted target value equals the observed target value. The Accuracy metric is the proportion of observations that are correctly classified.