

MSCA31010: Linear & Non-Linear Models

Winter Quarter 2023 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.xlsx**, we will train a generalized linear model to study how policy attributes affect Frequency. The generalized linear model has the following specifications.

- Response Variable: CLM_COUNT
- Distribution: Poisson
- Link Function: Natural logarithm
- Offset Variable: Natural logarithm of EXPOSURE
- Categorical Predictors: CAR_TYPE, CAR_USE, EDUCATION, GENDER, MSTATUS, PARENT1, RED_CAR, REVOKED, and URBANICITY. **Reorder the categories of each predictor in ascending order of the number of observations.**
- Interval Predictors: AGE, BLUEBOOK, CAR_AGE, HOME_VAL, HOMEKIDS, INCOME, YOJ, KIDSDRIV, MVR_PTS, TIF, and TRAVTIME. **Please divide BLUEBOOK, HOME_VAL, and INCOME by 1000 before training the model.**
- The model always includes the Intercept term.

We will use all complete observations with non-missing positive exposure for training all models. Therefore, we will drop all missing values (i.e., NaN) casewise in all the predictors and the target variable.

Question 1 (20 points)

We will first establish some baselines for reference. For instance, we will train the Intercept-only model. This model does not include any predictors except for the Intercept term.

- (10 points). Please generate a vertical bar chart to show the frequency of the number of claims.
- (10 points). What is the log-likelihood value, the Akaike Information Criterion (AIC) value, and the Bayesian Information Criterion (BIC) value of the Intercept-only model? Here are the formulas for AIC and BIC. Suppose l is the log-likelihood of a model, p is the number of non-aliased parameters in the model, and n is the number of observations used for training the model. Then $AIC = -2l + 2p$ and $BIC = -2l + p \log_e n$.

Question 2 (30 points)

Use the Forward Selection method to build our model. **The Entry Threshold is 0.01.**

- (10 points). Please provide a summary report of the Forward Selection in a table. 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.

- b) (10 points). Our final model is the model when the Forward Selection ends. What are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) of your final model?
- c) (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, the 95% asymptotic confidence intervals, and the exponentiated parameter estimates. Conventionally, aliased parameters have zero standard errors and confidence intervals.

Question 3 (30 points)

We will use accuracy metrics to assess the Intercept-only model and our final model in Question 2. These metrics inform us from various perspectives how well the predicted number of claims agrees with the observed number of claims.

- a) (10 points). Calculate the Root Mean Squared Error, the Relative Error, the Pearson correlation, the Distance correlation, and the R-squared metrics for the Intercept-only model.
- b) (10 points). Calculate the Root Mean Squared Error, the Relative Error, the Pearson correlation, the Distance correlation, and the R-squared metrics for our final model in Question 2.
- c) (10 points) We will compare the goodness-of-fit of your model with that of the saturated model. We will calculate the Pearson Chi-Squares and the Deviance Chi-Squares statistics, their degrees of freedom, and their significance values. Based on the results, do you think your model is statistically the same as the saturated Model?

Question 4 (20 points)

You will visually assess your final model in Question 1. Please color-code the markers according to the magnitude of the Exposure value. You must properly label the axes, add grid lines, and choose appropriate tick marks to receive full credit.

- a) (10 points). Plot the Pearson residuals versus the observed number of claims.
- b) (10 points). Plot the Deviance residuals versus the observed number of claims.