MSCA31010: Linear & Non-Linear Models

Winter Quarter 2023 Assignment 4

In insurance ratemaking, the term *Severity* is defined as the claim amount divided by the number of claims of a policy. Actuaries conventionally assume a Gamma distribution for Severity. Using the **claim_history.xlsx**, we will train a Gamma regression model to study how policy attributes affect Severity. The model has the following specifications.

- Response Variable: Severity = CLM_AMT / CLM_COUNT if CLM_COUNT > 0
- Distribution: Gamma
- Link Function: Natural logarithm
- Offset Variable: None
- 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 first drop all missing values casewisely in all the predictors and the target variable. Then, we will only use complete observations with a positive number of claims for training all models.

Question 1 (20 points)

We will first establish a baseline model for reference. To that end, we will train the Intercept-only model. This model does not include any predictors except for the Intercept term.

a) (10 points). Please generate a histogram and a horizontal boxplot to show the distribution of Severity. For the histogram, use a bin-width of $500 and put the number of policies on the vertical axis. Put the two graphs in the same chart where the histogram is above the boxplot.

b) (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?

Question 2 (30 points)

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

a) (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 Severity agrees with the observed Severity.

a) (10 points). Calculate the Root Mean Squared Error, the Relative Error, the Pearson correlation, the Distance correlation, and the Mean Absolute Proportion Error 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 Mean Absolute Proportion Error 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 2. 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 Severity.

b) (10 points). Plot the Deviance residuals versus the observed Severity.