MSCA31010: Linear & Non-Linear Models

Winter Quarter 2023 Assignment 5

Question 1 (50 points)

In insurance ratemaking, the term *Pure Premium* is defined as the claim amount divided by the Exposure duration. Nowadays, actuaries assume a Tweedie distribution for the claim amount. Using the **claim_history.xlsx**, we will train a Tweedie regression model to study how policy attributes affect the claim amount. The model has the following specifications.

- Response Variable: CLM_AMT
- Distribution: Tweedie
- 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 first drop all missing values casewisely in all the predictors, the target variable, and the offset variable. Then, we will only use complete observations for training our models.

a) (10 points). We will first estimate the Tweedie distribution's Power parameter p and Scale parameter ϕ. To this end, we will calculate the sample means and the sample variances of the claim amount for each value combination of the **categorical** predictors. Then, we will train a linear regression model to help us estimate the two parameters. What are their values? Please provide us with your appropriate chart.

b) (10 points). We will use the Forward Selection method to enter predictors into our model. Our entry threshold is 0.05. 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 Model Degree of Freedom (i.e., the number of non-aliased parameters), (4) the Quasi-Loglikelihood value, (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.

c) (10 points). We will calculate the Root Mean Squared Error, the Relative Error, the Pearson correlation, and the Distance correlation between the observed and the predicted claim amounts of your final model. Please comment on their values.

d) (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. Please also provide us with the final estimate of the Tweedie distribution's scale parameter ϕ.
e) (10 points). Please generate a Two-way Lift chart for comparing your final model with the Intercept-only model. Based on the chart, what will you conclude about your final model?

Question 2 (50 points)

Krall, Uthoff, and Harley (1975) analyzed data from a study on multiple myeloma in which researchers treated sixty-five patients with alkylating agents. Of those patients, forty-eight died during the study, and seventeen survived.

The data set is in the **myeloma.csv**. The variable **Time** represents the survival time in months from diagnosis. The variable **VStatus** consists of two values, 0 and 1, indicating whether the patient was alive or dead, respectively, at the end of the study. If the value of **VStatus** is 1, the patient died during the study. If the value of **VStatus** is 0, the patient was still alive at the end of the study and the corresponding value of Time is censored.

Reference: John M. Krall, Vincent A. Uthoff, and John B. Harley (1975). "A Step-Up Procedure for Selecting Variables Associated with Survival." *Biometrics*, volume 31, number 1, pages 49 – 57.

a) (10 points). How many risk sets are there?
b) (10 points). We will use the Kaplan-Meier Product Limit Estimator to create the life table. Please provide us with the life table.
c) (10 points). According to the life table, what is the Probability of Survival and the Cumulative Hazard at a survival time of 18 months? What do these two values mean to a layperson?
d) (10 points). Please generate the Survival Function graph using the Kaplan-Meier Product Limit Estimator life table. Since we measure the Time variable in the number of months, we will specify the x-axis ticks from 0 with an increment of 12. Besides plotting the Survival Function versus Time, you must also add the 95% Confidence Band. You might use the matplotlib `fill_between()` function to generate the Confidence Band as a band around the Survival Function. To receive the full credits, you must label the chart elements properly.
e) (10 points). Use Linear Interpolation to determine the Median Survival Time (in number of months) from the Kaplan-Meier Product Limit Estimator life table. Please round your answer up to the tenths place.