# MSCA31010: Linear & Non-Linear Models Winter Quarter 2023 Assignment 5

Question 1 (50 points)

1. (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.

Chart, scatter chart

Description automatically generated

**Power parameter 𝑝 : 1.2840608802234919**

**Scale parameter phi : 495.39032035257253**

1. (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.

Table

Description automatically generated

1. (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.

**RMSE : 4116.064009419275**

**Relative Error : 1.0078985075249884**

**Pearson Correlation : 0.18768098705727354**

**Distance Correlation : 0.27019055675583464**

A lower RMSE indicates better performance of the model. In this case, the RMSE value of 4116.06 suggests that there is a significant difference between the predicted and actual values. The relative error value of 1.0079 indicates that the predicted values are slightly higher than the actual values. The Pearson correlation value of 0.1876 suggests that there is a weak positive correlation between the predicted and actual values. The distance correlation value of 0.2702 indicates a weak dependence between the predicted and actual values**. Overall, these values suggest that the predictive model may not be performing very well. The RMSE value is quite high, and the correlation values suggest weak relationships between the predicted and actual values.**

1. (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.

Table

Description automatically generated

Final estimate of Tweedie distribution’s scale parameter 𝜙 = **319.3817800403747**

1. (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?

Chart, line chart

Description automatically generated

From the chart we can infer that, our final model performs best when we target the top 10% of the population (decile of Impact = 1) but gradually decreases as the target population increases. Our model performance is very poor as we can see that post decile of impact = 7 our model performance is worse that our baseline model, i.e., the Intercept only model.

Question 2 (50 points)

1. (10 points). How many risk sets are there?

A risk set is defined as the set of individuals who are at risk of experiencing an event at a given time point. For example, in our case, the risk set at a particular time point consists of all patients who have not yet died in our multiple myeloma study.

From the life table we can identify that the number of risk sets is **38**

1. (10 points). We will use the Kaplan-Meier Product Limit Estimator to create the life table. Please provide us with the life table.

**Table

Description automatically generated**

1. (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?

**Survival Time : 18 months**

**Probability of Survival : 0.509496**

**Cumulative Hazard : 0.653547**

These values tell us that tell us that the probability of surviving up to 18 months is **50.95%**, and the cumulative probability of experiencing the event of interest (death due to multiple myeloma in our case) up to this time point (i.e., 18 months) is **65.35%**

1. (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.

**Chart, line chart

Description automatically generatedChart

Description automatically generated**

1. (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.

Median Survival Time (in number of months) is **18.2 months**.